REPORT DOCUMENTATION PAGE

OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	5/9/97	FINAL TECHNICAL REPORT (3/1/94 - 2/28/97)

4. TITLE AND SUBTITLE

TITLE: Dynamic Networks Techniques for Autonomous Planning and Control

SUBTITLE: Probabilistic Counterfactuals

5. FUNDING NUMBERS

G - F49620-94-1-0173

Rec'd Copy 10/14/97

6. AUTHOR(S)

Professor Judea Pearl

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

UCLA
Computer Science Department
4532 Boelter Hall
Los Angeles, CA 90095-1596

8. PERFORMING ORGANIZATION REPORT NUMBER

932213-00-A03

442510-22540

AFOSR-TR-97-

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Air Force / Office of Scientific Research
110 Duncan Avenue, Suite B115
Bolling Air Force Base
Washington, DC 20332-0001

nm

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

G6411

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

We have reformulated Bayesian networks as carriers of causal information. Teh result is a more natural understanding of what the networks stand for, what judgments are required in constructing the network and, most importantly, how actions and plans are to be handled within the framework of standard probability theory. Starting with functional description of physical mechanisms, we were able to derive the standard probabilistic properties of Bayesian networks and to show:

* how the effects of unanticipated actions can be predicted from the network topology,
* how qualitative causal judgments can be integrated with statistical data,
* how actions interact with observations, and
* how counterfactuals sentences can be formulated and evaluated.

14. SUBJECT TERMS

Keywords: Causation, counterfactuals, Bayesian networks

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
unclassifed	unclassified	unclassified	SAR

DYNAMIC NETWORKS TECHNIQUES FOR AUTONOMOUS PLANNING AND CONTROL

Professor Judea Pearl
Principal Investigator
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024-1596, USA
judea@cs.ucla.edu

FINAL TECHNICAL REPORT: 3/1/94–5/31/97
AWARD NO: F49620-94-1-0173

Objectives:

- Real-time planning under uncertainty using qualitative approximations of probabilities and utilities.
- Learning network structures from empirical data, including the identification of hidden causes and stable mechanisms, by local analysis.

Status of effort:

To facilitate the construction of practical decision making systems, we have focused our research effort toward basing Bayesian networks directly on causal relationships. The result is a more natural understanding of what the networks stand for, what judgments are required in constructing the network and, most importantly, how actions and plans are to be handled within the framework of standard probability theory. Starting with functional description of physical mechanisms. we were able to derive the standard probabilistic properties of Bayesian networks and to show, additionally:

- how the effects of unanticipated actions can be predicted from the network topology,
- how qualitative causal judgments can be integrated with statistical data,
- how actions interact with observations, and
- how counterfactuals sentences can be formulated and evaluated,

Additionally, we have established an axiomatic characterization of causal dependencies, analogously to the graphoid characterization of informational dependencies. Finally, we have demonstrated that network-learning techniques, in the presence of hidden variables, have enormous scope of new applications, ranging from skill acquisition by autonomous agents, to the analysis of treatment effectiveness in clinical trials.

Our research has shown the feasibility of predicting the merit of an action or a plan by observing the performance of other agents, for example, watching the sequence of requests provided by a user of a computer system or the sequence of actions taken by a skilled operator of a complex system.

Accomplishments/New Findings:

The following specific results were obtained during the period of performance:

- Graphical criteria were developed for identifying conditional independence relationships induced by systems with feedback (Pearl & Dechter 1996).
- Computer programs were developed to assist clinicians with assessing the efficacy of treatments in experimental studies for which subject compliance is imperfect (Chickering & Pearl 1996).

- Axiomatic characterization was given for causal-relevance relationships of the form: "Changing X will not affect Y if we hold Z constant" (Galles & Pearl 1996).
- The notion of "identification" was extended to non-parametric systems (Pearl 1995b) and techniques were developed for non-parametric identification of cause-effect relationships from nonexperimental data (Pearl 1995a; Pearl & Robins 1995; Balke & Pearl 1995; Galles & Pearl 1995).
- Deriving algebraic expressions for identifiable causal effects (both total and direct) in non-parametric structural models with latent variables.
- Selecting sufficient set of measurements (covariates or confounders) that permit unbiased estimation of causal effects in observational studies.
- Predicting (or bounding) treatment effectiveness from trials with imperfect compliance.
- Estimating (or bounding) counterfactual probabilities from statistical data (e.g., John, who was treated and died, would have had 90% chance of survival had he not been treated)
- A formal model has been developed, based on dynamic structural equations, which generalizes and unifies the structural and counterfactual approaches to causal inference, explicates their conceptual and mathematical bases and resolves their technical difficulties. A simple rule was devised for translating a problem back and forth, between the structural and counterfactual representations, and choose the one most appropriate for analysis.
- It has been proven that the structural and counterfactual formalisms are equivalent in recursive causal models (i.e., systems without feedback) but not when feedback is considered possible.

Personnel Supported:

Principal Investigator:

Judea Pearl

Post-Docs:

Adnan Darwiche

Graduate Students:

Alex Balke (PhD, 1995), "Probabilistic Counterfactuals: Semantics, Computation, and Applications"

David Chickering (PhD, 1996), "Learning Bayesian Networks from Data"

David Galles (PhD, expected June 1997), "Causal Theories: A Formalism for Modeling Action and Intervention"

Huy Cao

Kevin Chang

Research Associates:

Rina Dechter

Avi Dechter

Norman Dalkey

Publications (3/1/94–2/28/97):

Pearl, J., "From Imaging and Stochastic Control to a Calculus of Actions," *Symposium Notes of the 1994 AAAI Spring Symposium on Decision-Theoretic Planning*, Stanford, CA, 204-209, March 21-23, 1994.

Pearl, J., "From Adams' conditionals to default expressions, causal conditionals, and counterfactuals," in E. Eells and B. Skyrms (Eds.), *Probability and Conditionals*, Cambridge University Press, New York, NY, 47-74, 1994.

Pearl, J. and N. Wermuth, "When Can Association Graphs Admit A Causal Explanation?," in P. Cheeseman and W. Oldford (Eds.), *Selecting Models and Data, Artificial Intelligence and Statistics IV*, Springer-Verlag, 205-214, 1994.

Darwiche, A. and J. Pearl, "On the Logic of Iterated Belief Revision," in R. Fagin (Ed.), *Proceedings of the 1994 Conference on Theoretical Aspects of Reasoning about Knowledge (TARK '94)*, Pacific Grove, CA, 5-23, Mar. 1994. To appear in *Artificial Intelligence*, Spring 1997.

Darwiche, A. and J. Pearl, "Symbolic Causal Networks for Planning under Uncertainty," In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, WA, Volume I, 238-244, 1994.

Tan, S-W., "Qualitative Decision Theory," In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, WA, Volume II, 928-933, July 31 - August 4, 1994.

Tan, S-W. and J. Pearl, "Specification and Evaluation of Preferences for Planning under Uncertainty," in J. Doyle, E. Sandewall, and P. Torasso (Eds.), *Proceedings of the Fourth International Conference, Principles of Knowledge Representation and Reasoning (KR-94)*, Bonn, Germany, Morgan Kaufmann, San Francisco, CA, 530-539, May 1994.

Pearl, J., "A Probabilistic Calculus of Actions," In R. Lopez de Mantaras and D. Poole (Eds.), *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence (UAI-94)*, Morgan Kaufman, San Mateo, CA, 454-462, 1994.

Balke, A., and Pearl, J., "Counterfactual Probabilities: Computational Methods, Bounds, and Applications," in R. Lopez de Mantaras and D. Poole (Eds.), *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI-94)*, Morgan Kaufmann, San Mateo, CA, 46-54, July 29-31, 1994.

Tan, S-W. and Pearl, J., "Exceptional Subclasses in Qualitative Probability," in R. Lopez de Mantaras and D. Poole (Eds.), *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence (UAI-94)*, Morgan Kaufmann, San Mateo, CA, 553-559, July 29-31, 1994.

Balke, A. and Pearl, J., "Probabilistic Evaluation of Counterfactual Queries," in *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, WA, Volume I, 230-237, July 31 - August 4, 1994.

Pearl, J. and Verma, T., "A Theory of Inferred Causation," in D. Prawitz, B. Skyrms and D. Westertahl (Eds.), *Logic, Methodology and Philosophy of Science IX*, Elsevier Science B.V., 789-811, 1994.

Balke, A. and Pearl, J., "Universal Formulas for Treatment Effects from Noncompliance Data," in N.P. Jewell, A.C. Kimber, M.-L.T. Lee, and G.A. Whitmore (Eds.), *Lifetime Data: Models in Reliability and Survival Analysis*, Kluwer Academic Publishers, Dordrecht, 39–43, 1995.

Pearl, J., "Causal Inference from Indirect Experiments," *Artificial Intelligence in Medicine*, Vol. 7, No. 6, 561–582, 1995.

Pearl, J., "On the Testability of Causal Models with Latent and Instrumental Variables," in P. Besnard and S. Hanks (Eds.), *Uncertainty in Artificial Intelligence 11*, Morgan Kaufmann, San Francisco, CA, 435–443, 1995.

Pearl, J. and Robins, J., "Probabilistic evaluation of sequential plans from causal models with hidden variables," in P. Besnard and S. Hanks (Eds.), *Uncertainty in Artificial Intelligence 11*, Morgan Kaufmann, San Francisco, CA, 444–453, 1995.

Pearl, J., "Causation, Action, and Counterfactuals," in A. Gammerman (Ed.), *Computational Learning and Probabilistic Reasoning*, John Wiley and Sons, New York, Chapter 6, 235–255, 1995.

Tan, S-W. and Pearl, J., "Specificity and Inheritance in Default Reasoning," *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (IIJCAI-95), Montreal, Quebec, August 20-25, Vol. 2, 1480–1486, 1995.

Pearl, J., "From Bayesian Networks to Causal Networks," in A. Gammerman (Ed.), *Bayesian Networks and Probabilistic Reasoning*, Alfred Walter Ltd., London, 1-31, 1994. In *Proceedings of the UNICOM Seminar on Adaptive Computing and Information Processing*, Brunel University, London, pp. 165-194, January 25-27, 1994. Also in G. Coletti, D. Dubois, and R. Scozzafava (Eds.), *Mathematical Models for Handling Partial Knowledge in Artificial Intelligence*, Plenum Publishing, New York, NY, 1995.

Pearl, J., "Causation, Action, and Counterfactuals," Extended Abstract in A. Cohn (Ed), *11th European Conference on Artificial Intelligence (ECAI-94)*, John Wiley and Sons, Ltd., 826–828, 1994.

Galles, D. and Pearl, J., "Testing Identifiability of Causal Effects," In P. Besnard and S. Hanks (Eds.), *Uncertainty in Artificial Intelligence 11*, Morgan Kaufmann, San Francisco, CA, 185–195, 1995.

Balke, A. and Pearl, J., "Counterfactuals and Policy Analysis in Structural Models," in P. Besnard and S. Hanks (Eds.), *Uncertainty in Artificial Intelligence 11*, Morgan Kaufmann, San Francisco, CA, 11–18, 1995.

Pearl, J., "Bayesian Networks," in M. Arbib (Ed.), *Handbook of Brain Theory and Neural Networks*, MIT Press, 149–153, 1995.

Pearl, J., "Causal diagrams for empirical research, (with discussion)," *Biometrika*, 82(4), 669–710, December 1995.

Goldszmidt, M. and Pearl, J., "Qualitative Probabilities for Default Reasoning, Belief Revision and Causal Modeling," *Artificial Intelligence*, Vol. 84, No. 1-2, 57–112, 1996.

Ben-Eliyahu, R. & Dechter, R., "Default Reasoning Using Classical Logic," *Artificial Intelligence Journal*, Volume 84, Issue 1-2, 113–150, July 1996.

Chickering, D.M. and Pearl, J., "A Clinician's Apprentice for Analyzing Non-compliance," in *Proceedings of the National Conference on Artificial Intelligence (AAAI-96)*, Portland, OR, 1269-1276, August 1996.

Pearl, J. and Dechter, R., "Identifying independencies in causal graphs with feedback," in *Proceedings of Uncertainty in Artificial Intelligence (UAI-96)*, Portland, OR, 240-246, August 1996.

Pearl, J. and Dechter, R., "Identifying independencies in causal graphs with feedback," In E. Horvitz and E. F. Jensen (Eds.), *Uncertainty in Artificiall Intelligence, Proceedings of the Twelfth Conference*, Morgan Kaufmann: San Francisco, CA, 240–246, August 1996.

Pearl, J., "Graphical Models for Probabilistic and Causal Reasoning," In Allen B. Tucker, Jr. (Ed.), *The Computer Science and Engineering Handbook*, Chapter 31, CRC Press, Inc., 697–714, 1997.

Pearl, J., "Structural and Probabilistic Causality," in D.R. Shanks, K.J. Holyoak, and D.L. Medin (Eds.), *The Psychology of Learning and Motivation, Vol. 34: Causal Learning.* Academic Press, San Diego, CA, 393–435, 1996.

Pearl, J. and Goldszmidt, M., "Probabilistic Foundations of Reasoning with Conditionals," in G. Brewka (Ed.), *Principles of Knowledge Representation*, CSLI Publications, 33–68, 1996.

Pearl, J., "Causation, Action, and Counterfactuals," in Yoav Shoham (Ed.), *Theoretical Aspects of Rationality and Knowledge, Proceedings of the Sixth Conference (TARK 1996)*, The Netherlands, 51–73, March 17-20, 1996

Pearl, J., "Decision Making Under Uncertainty," Prepared for *CRC Handbook* chapter for special 50th-anniversary issue of Computing Surveys, 1996.

Meiri, I., "Combining Qualitative and Quantitative Constraints in Temporal Reasoning," *Artificial Intelligence*, 87, 1–46, 1996.

Dechter, R. & Dechter, A., "Structure-Driven Algorithms for Truth Maintenance," *Artificial Intelligence*, 82, 1–20, 1996.

Paz, A., Pearl, J., & Ur, S., "A New Characterization of Graphs Based on Interception," *Journal of Graph Theory*, Vol. 22, No. 2, 125–136, 1996.

Pearl, J., "The Art and Science of Cause and Effect," Given October 29, 1996 as part of the UCLA 81st Faculty Research Lecture Series.

Pearl, J., "TETRAD and SEM," UCLA Computer Science Department, Technical Report (R-244), June 1996. Commentary on "The TETRAD Project: Constraint Based Aids to Causal Model Specification" by R. Scheines, P. Spirtes, C. Glymour, C. Meek, and T. Richardson. Prepared for *Multivariate Behavioral Research.*

Galles, D. & Pearl, J., "Axioms of Causal Relevance," Preliminary version in *Proceedings of the Fourth International Conference on Mathematics and AI*, Fort Lauderdale, FL, 64–67, January, 1996. Revision I submitted to *Artificial Intelligence*, November 1996.

Pearl, J., "Bayesian Networks," UCLA Computer Science Department, Technical Report (R-246), November 1996. To appear in *MIT Encyclopedia of the Cognitive Sciences.*

Pearl, J., "Comments on R.W. Olfords' 'A Physical Device for Demonstrating Confounding, Blocking, and the Role of Randomization in Uncovering a Causal Relationship'," *The American Statistician*, Vol. 50, No. 4, 387–388, November 1996.

Pearl, J., "Graphs, Structural Models and Causality," UCLA Computer Science Department, Technical Report (R-247), December 1996. (A condensed version of this paper has appeared in *Biometrika*, 82(4), 669–710, December 1995 under the title "Causal Diagrams for Experimental Research". Prepared for AAAI/MIT Press volume of *Causation, Bayes Networks, and Machine Discovery.*

Balke, A. & Pearl, J., "Nonparametric Bounds on Causal Effects from Partial Compliance Data," UCLA Computer Science Department, Technical Report (R-199), Revision II, November 1996. To appear in *JASA*, September 1997.

Darwiche, A. & Pearl, J., "On the Logic of Iterated Belief Revision," *Artificial Intelligence*, Vol. 89, Nos. 1-2, 1–29, January 1997.

Pearl, J., "The New Challenge: From a Century of Statistics to an Age of Causation," *Proceedings the IASC Second World Congress*, Pasadena, CA, February 1997.

Pearl, J., "On the Identification of Nonparametric Structural Models," in M. Berkane (Ed.), *Latent Variable Modeling with Application to Causality Conference*, Springer-Verlag, Lecture Notes in Statistics, 29–68, 1997.

Pearl, J., "Causation, Action, and Counterfactuals," In M.L. Dalla Chiara et al. (Eds.), *Logic and Scientific Methods*, Kluwer Academic Publishers, Netherlands, 355–375, 1997.

Interactions/Transitions:

1994 4th International Conference, Principles of Knowledge Representation & Reasoning (KR-94)
1994 International Research Conference on Lifetime Data Models in Reliability & Survival Analysis

1994 Conference on Theoretical Aspects of Reasoning about Knowledge (TARK-94)
1994 AAAI Spring Symposium on Decision-Theoretic Planning
1994 12th National Conference on Artificial Intelligence (AAAI-94)
1994 Latent Variable Modelling with Application to Causality Conference
1994 10th Conference on Uncertainty in Artificial Intelligence (UAI-94)
1994 11th European Conference on Artificial Intelligence (ECAI)
1995 11th Conference on Uncertainty in Artificial Intelligence (UAI-95)
1995 14th International Joint Conference on Artificial Intelligence (IJCAI-95)
1995 Annual Meeting of the National Academy of Engineering
1996 International Symposium on AI/Math
1996 6th Conference on Theoretical Aspects of Rationality and Knowledge (TARK-96)
1996 National Conference on Artificial Intelligence (AAAI-96)
1996 12th Conference on Uncertainty in Artificial Intelligence (UAI-96)

New discoveries, inventions, or patent disclosures:

None.

Honors/Awards:

Professor Pearl is a recipient of the RCA Laboratories Achievement Award (1965), and a NATO Senior Fellowship in Science (1975). He is a Fellow of IEEE, a founding Fellow of AAAI, and a member of the National Academy of Engineering (NA E). Recently named UCLA's 81st Faculty Research Lecturer.

UNIVERSITY OF CALIFORNIA

Los Angeles

Structural Causal Models: A Formalism for Reasoning About Actions and Counterfactuals

A dissertation submitted in partial satisfaction

of the requirements for the degree

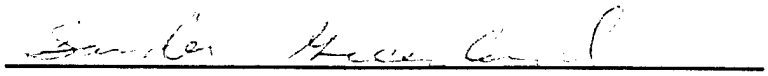
Doctor of Philosophy in Computer Science

by

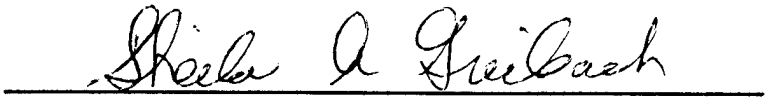
David Jerome Galles

1997

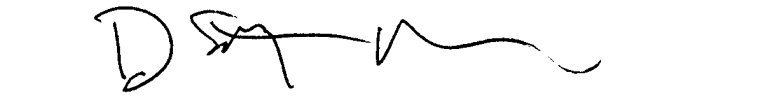
The dissertation of David Jerome Galles is approved.



Sander Greenland



Sheila A. Greibach



D. Stott Parker

Judea Pearl, Committee Chair

University of California, Los Angeles

1997

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisor, Judea Pearl. I was blessed to be able to work with a researcher of his caliber. The members of my committee, D. Stott Parker, Sheila Greibach, and Sander Greenland, deserve thanks for valuable insights that added greatly to the quality of this document. I would also like to thank Kaoru Mulvihill for frequently coming to my aid as deadlines loomed. Along with every graduate of the UCLA Computer Science Department, I am indebted to Verra Morgan for ensuring that I jumped through all of the requisite hoops along the way to graduation. Finally, my deepest gratitude goes to my wife, Julie. Without her constant support and encouragement, I would never have finished this degree.

VITA

1969	Born, Walnut Creek, California
1989–1992	Section Leader, Computer Science Department, Stanford University, Palo Alto, California.
1992	B.S. (Computer Science), with Distinction, Stanford University, Palo Alto, California.
1993–1995	Teaching Assistant, Computer Science Department, University of California, Los Angeles.
1994	M.S. (Computer Science), University of California, Los Angeles.
1995-1997	Research Assistant, Computer Science Department, University of California, Los Angeles.
1995-1997	Instructor, Loyola-Marymount site, Center for Talented Youth, Johns Hopkins University.

ABSTRACT OF THE DISSERTATION

Structural Causal Models: A Formalism for Reasoning About Actions and Counterfactuals

by

David Jerome Galles

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 1997

Professor Judea Pearl, Chair

Our everyday language is permeated with causal utterances. In-depth understanding of an event or action is associated with comprehension of the causal mechanisms that led to the event or governed the action. This dissertation gives formal underpinnings to some of the intuitive notions of causation, in the language of structural causal models. This formalism, which is tailored after the structural equation model of engineering and economics, provides a language for specifying the precise meaning of concepts such as influence, causal relevance, counterfactuals, probability of causes and probability of effects. We model actions as local modifications of causal models, that is, the replacement of a set of equations with other equations.

We sharpen our understanding of causality by developing an axiomization for causal relevance in causal models. We examine two versions of conditional causal irrelevance: probabilistic, "If we hold Z constant, changing X cannot change the

probability of Y" and deterministic, "If we hold Z constant, changing X cannot alter the value of Y in any circumstance." Comparison of the behavior of these two types of relevance to the behavior of conditional independence reveals some of the fundamental differences between causation and correlation.

Finally, we demonstrate how causal models can be used to calculate quantities of interest. We provide a polynomial-time algorithm that decides, given the graph associated with a causal model, whether the causal effect of one variable on another can be determined from data obtained under controlled conditions. Whenever such a determination is feasible, then the algorithm yields a closed-form expression for the causal effect.

CHAPTER 1

Causality

1.1 Introduction

Causation is ubiquitous in everyday language. For instance, when asked why there is a stain on the carpet, one might reply "I knocked over my coffee mug." This is a causal explanation of the event—it implies a chain of events, each of which is the direct cause of the next: knocking over the mug caused the coffee in the mug to spill, and the coffee that spilled on the floor stained the carpet. On any given day, we make dozens of such explanations.

Given the prevalence of causality and causal language in everyday language, we would expect scientific discourse to be laced with causal concepts. However, when we look at statistics, which is the formal language commonly used by scientists, we find a notable absence of causal concepts. The same is true for standard first-order logic, where there is no distinction between causation and implication. Why is causality, which we find so useful in everyday life, missing from the formal languages of statistics and logic?

One is the lack of rigorous, mathematical, and useful definitions of causal terms. This dissertation seeks to offer rigorous definitions of causal terms such as influence, causal irrelevance, direct effect. Of course, on their own definitions are of little use. The point of crafting these definitions, and ensuring that they have

a rigorous mathematical basis, is to manipulate them. Giving causal language precise meaning will allow us to develop a calculus for answering queries about causation and to create useful tools for policy analysis as well as world modeling.

The formal language that we will use to obtain a specification of causality is that of *structural causal models*, which will be crafted and interpreted after the structual equation models used in engineering and statistics. Thoughout this thesis, we will use the abbreviations *structural model* qnd *causal model* to emphasize different aspects of structural causal models. Causal models offer a succinct language for discussing the effects of actions in the world. As such, they are a useful tool for modeling interventions in a wide range of applications. After developing the theoretical foundations of causal models, this dissertation will provide a mathematical characterization of causal relevance, that is akin to the work in graphoids on observational relevance. Various useful extensions to causal models which allow for time-varying as well as steady-state systems will be investigated. Finally, practical applications of causal models will be demonstrated.

1.2 Previous Work

When developing a formalism for describing and reasoning about causality, researchers often present the concept of causality within the framework of causal counterfactuals. A counterfactual statement takes the form, "If A were true, then B would be true as well." Much of the work in evaluating these counterfactuals involves considering the minimal change needed to make A true and testing whether B is then true as well. The tricky part is defining what exactly is meant by "minimal change."

1.2.1 Lewis's Counterfactuals

Lewis approaches the concept of a minimal change through the concept of a *closest world* [Lew73a]. Lewis defines the counterfactual statement "If it were the case that A, then it would be the case that B," written $A \boxdot\!\!\rightarrow B$, as true at world i iff some accessible AB-world is closer to i than any $A\overline{B}$ world, if there are any A-worlds accessible from i. Thus, given any world i, we can imagine spheres of similarity around i. A set of worlds S is a sphere around i iff every S-world is accessible from i and is closer to i than any world not in S. We call a sphere S *A-permitting* if there exists some world in S such that A holds. Thus, $A \boxdot\!\!\rightarrow B$ holds at i iff there exists some A-permitting sphere S such that $A \rightarrow B$ in every member of S. This is shown graphically in Figure 1.1.

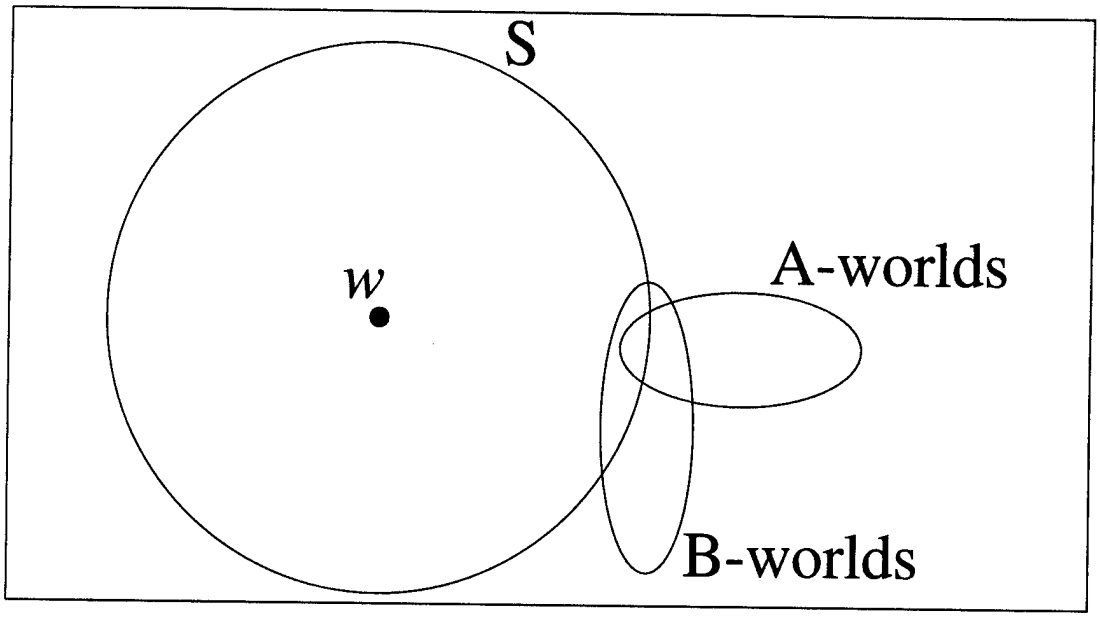


Figure 1.1: Graphical representation of $A \boxdot\!\!\rightarrow B$ in Lewis's formalism

To completely define a system using this formalism, we need to give definitions for accessibility and distance. On accessibility, Lewis keeps an open mind.

He cannot imagine a convincing case showing the need for inaccessibility but leaves accessibility in the formalism, arguing that accessibility restrictions can be dropped by making all worlds equally accessible. He purposely leaves out specification of possible distance measures in an effort to make the formalism as general as possible. He argues that because counterfactuals themselves are imprecise, it makes more sense to be precise about the link between two imprecise notions than it does to tie down an imprecise notion with precise one.

In essence, causal models as discussed in this thesis represent a commitment to a specific type of distance metric (i.e., "closest") in Lewis, one that is amiable to computer representation. While in Lewis's system, the distance metric is taken as primitive, in our system the distance metric is derived from a more fundamental notion of *mechanisms*. Our use of causal mechanisms as a primitive, rather than a more abstract distance measure, allows for easier manipulation of causal sentences by digital computers. We will show that, in many cases, this specific choice of the distance metric does not affect the set of counterfactual statements that we consider valid.

1.2.2 Iwasaki and Simon's Causality in Device Behavior

Open any physics textbook and you will find the physical world described with systems of simultaneous equations. Thus, such systems of equations would seem to be a natural vehicle for expressing causation. Iwasaki and Simon [IS86] have devised a method that imparts a causal flavor to systems of equations. To the sets of equations, they infer a *causal ordering* that tells which variables causally affect other variables in the system. Much of their work revolves around determining what the causal order should be, given a set of equations, each describing a

physical law.

In contrast, we take causal ordering as given and explore the utilization of the ordered system as a means for interpreting causal expressions, and predicting the effect of actions. Additionally, we incorporate uncertainty into the model which enables us to work with partially specified systems.

1.2.3 Balke's Probabilistic Counterfactuals

Balke [Bal95] investigates the relation between causal counterfactuals and causal models, and utilizes causal models to compute causal counterfactual queries. In this formalism, causal relations between variables are mediated by response-function variables.

Consider, for instance, a simple system of two binary variables, A and B, such that A has some causal influence on B, and B has no causal influence on A. There are four possible deterministic functions between A and B: B always has the value b_0 (regardless of the value of A); B has the value b_0 if A has the value a_0, and the value b_1 otherwise; B has the value b_1 if A has the value a_0, and the value b_0 otherwise; B has the value b_1 (regardless of the value of A). We can imagine a response variable r_b that dictates which of these deterministic functions describe the causal effect of A on B. Thus, the value of B becomes a deterministic function of A and the response-function variable: $b = f_b(a, r_b)$. Since there are four possible functional mappings from A to B, the response-function variable for B has a four-valued domain, $r_b \in \{0, 1, 2, 3\}$. The complete functional specification for B is

$$b = f_b(a, r_b) = h_{b,r_b}(a) \tag{1.1}$$

where each response function h_{b,r_b} can be specified as follows:

$$h_{b,0}(a) = b_0 \tag{1.2}$$

$$h_{b,1}(a) = \begin{cases} b_0 & \text{if } a = a_0 \\ b_1 & \text{if } a = a_1 \end{cases} \tag{1.3}$$

$$h_{b,2}(a) = \begin{cases} b_1 & \text{if } a = a_0 \\ b_0 & \text{if } a = a_1 \end{cases} \tag{1.4}$$

$$h_{b,3}(a) = b_1 \tag{1.5}$$

A probability distribution over r_b completes the parameterization of the model. Heckerman and Schachter [HS94] used a similar model in their analysis of counterfactuals

In Balke's approach, answering counterfactual queries such as "Given that we observe the set of variables C to have the value c, what is the probability the B would be equal to b, if A were a" is done in the following fashion: (1) start with some distribution over the response-function variables; (2) use observations of variables in C to obtain updated values for the response variables via the Bayesian updating techniques of [Pea88]; (3) modify the response function for A so that A always has the value a; (4) using the updated response functions, calculate a probability distribution for B.

1.2.4 Neyman-Rubin's Counterfactuals

Balke's response functions, as well as the counterfactual interpretation given in this dissertation, have close connections to the notion of "unit potential response" which has been used in statistics [Ney23, Rub74] as the basis for analyzing the effect of treatments on populations. The unit potential response, denoted $Y_x(u)$, stands for the counterfactual sentence "The value that Y would take in person u had X been x," where X stands for a type of treatment that a person can receive.

The essential difference between Neyman-Rubin' framework and the one explored in this dissertation is that the former takes $Y_x(u)$ as primitive while we treat it as a derived quantity, computed from the fundamental of the processes responsible for Y taking on the value $Y_x(u)$ as X changes to x. We treat u not as merely the index of an individual but, rather, as the set of attributes u that characterize the individual, the experimental conditions under study, and so on. In fact, every structural causal model can be translated into a set of counterfactual statements of the type used in the statistical literature [Pea95a, p. 703], and conversely, every counterfactual sentence can be treated as a constraint over the behavior of a structural causal model. Using our process-based semantics, however, uncovers properties of $Y_x(u)$ that were not formalized in the statistical literature.

1.3 Contributions

This dissertation gives a formal underpinning to some of the intuitive notions of causation by using the language of structural models to specify the precise meaning of concepts such as causal relevance and causal influence. The principle contributions of this dissertation consist of:

- Construction of a set of axioms for causal counterfactuals that is sound and complete relative to the formal interpretation of action and change.
- Implementation of notions of past state and time within the framework of causal models.
- Formal mathematical specification of two types of causal irrelevance: probabilistic and deterministic.

- Construction of a set of axioms for causal irrelevance that is similar to the set of graphoid axioms for informational irrelevance.

- A method of using graphs as theorem provers to validate properties of causal irrelevance.

- A polynomial-time algorithm for deciding the identifiability of control queries given the underlying graph of a causal model.

1.4 Overview

In Chapter 2, we give a formal description of causal models, along with examples and motivations. We also prove some properties of causal models, compare our formalism to Lewis's, and translate linguistic notions of causality into the language of structural models. In Chapter 3, we explore ways of adding the concept of memory to causal models. Our extensions allow for consideration of past state and time within this formalism. Chapter 4 sharpens our understanding of causality within the framework of structural models. We provide probabilistic and deterministic definitions of causal relevance and develop some axioms over each type of irrelevance. In Chapter 5, we demonstrate how causal models can be used to calculate quantities of interest. We take as our example the problem of identifiability and, ultimately, present a polynomial-time algorithm that determines the identifiability of a control query by utilizing the graph of a causal model. We summarize and make concluding remarks in Chapter 6.

CHAPTER 2

Causal Models

2.1 Introduction

This chapter gives a formal definition of causal models, which are based on structural equation models in engineering and economics [Haa43, Gol73, Sim70]. Unlike structural equation models, which assume linear interactions, causal models allow for each variable to be an arbitrary function of other variables in the model. Thus, any possible steady-state interaction can be modeled using a causal model. We build our definition of causal models in two stages. First, we define an *equation set.* We then refine the definition of an equation set to define a causal model (see [Pea95a]). Throughout this dissertation, we will refer to the terms causal model and structural model interchangeably.

2.2 Equation Set

An *equation set* is a representation of the causal mechanisms that govern a set of variables. The variable set is broken into two categories, exogenous variables and endogenous variables. Each endogenous variable receives its value from a deterministic causal mechanism. This causal mechanism is represented by a deterministic function. Thus, for each endogenous variable X, there will be a function f_X that completely determines the value of X, given all the other variables in the

system. Formally, an equation set is defined as follows:

Definition 1 (equation set) *An* equation set *is a 3-tuple*

$$M = \langle V, U, F \rangle$$

where

(i) $V = \{X_1, \ldots, X_n\}$ *is a set of endogenous variables determined within the system,*

(ii) $U = \{U_1, \ldots, U_m\}$ *is a set of exogenous variables that represent disturbances, abnormalities, assumptions, or boundary conditions, and*

(iii) $F = \{f_{X_i}\}$ *is a set of n deterministic, nontrivial functions, each of the form*

$$x_i = f_{X_i}(\mathbf{pa}_i, u) \quad i = 1, \ldots, n \tag{2.1}$$

where $\mathbf{pa}_i$ *are the values of a set of variables* $PA_i \subseteq V \setminus X_i$ *(connoting parents), called the* direct causes *of* X_i.

Each equation $x_i = f_{X_i}$ *has a privileged variable* X_i*, which appears only once in the equation, as the only variable to the left of the equals sign.*

It is important to note that each equation $x_i = f_{X_i}$ has a privileged variable X_i. The function f_{X_i} represents the causal mechanism that determines the value of X_i. Thus, if we take an equation set that contains the variables X_i and X_j, and replace the equation f_{X_i} with a linear combination of the equations f_{X_i} and f_{X_j}, we would no longer have an equation set, because each equation would no longer represent the mechanism that affects the value of a single privileged variable. Thus, any group of equations is not an equation set.

The restriction that the equations f_{X_i} be nontrivial functions of the parents $\mathbf{pa}_i$ means that for each variable X_n in $\mathbf{pa}_i$, there exists two values x_n and x'_n such that

$$f_{X_i}(pa'_i, x_n) \neq f_{X_i}(pa'_i, x_n)$$

where pa'_i is a set of values for $\mathbf{pa}_i \setminus X_n$.

Thus, $f_{X_1}(x_2, x_3) = x_2$ is a nontrivial function of X_2 but a trivial function of X_3. This restriction does not so much prohibit which variables can be parents as define what a parent is.

2.3 Valid Causal Models

Now that we have a definition of an equation set, we can use it to define a causal model. A *causal model* is an equation set that meets two restrictions on the equations f_{X_i}:

Definition 2 (causal model) *A* casual model *is an equation set that meets the following two restrictions:*

(i) The set of equations $\{f_{X_i}\}$ has a unique solution for $X_1, \ldots, X_n$ given any value of the disturbances $U_1, \ldots, U_m$.

(ii) If we replace any subset of the set of equations $\{f_{X_i}\}$ with constant functions $f_{X_j} = c$ (where c is any constant), the remaining equations will also have a unique solution for any value of the disturbances $U_1, \ldots, U_m$.

Examples of causal models can be found in Section 2.6.

The uniqueness assumption is equivalent to the requirement that F represents a deterministic physical system in equilibrium. Assuming that all relevant

boundary conditions U are accounted for, such a system can only be in one state.

The assumption that there is a unique solution for $X_1, \ldots, X_n$ imposes some restrictions on the functions f_{X_i}. At first, this assumption sees overly restrictive. Why restrict the equation set to having a unique fixed point for all possible values of the exogenous variables? Doesn't this needlessly restrict what we can express using causal models? Upon further inspection, however, this requirement is not as restrictive as it first appears. Consider an equation set with two fixed points for a particular value of U. If we consider the exogenous variables to be the world state as defined in physics, we have a world state that includes a variable whose value is not determined. Thus, we have not completely specified the world state and do not have a complete model until we add some more information to the world state. If we wish to describe a world state in which the value of a variable is only stochastically determined, we need to add an additional variable U_{m+1} that governs the stochastic interaction to U.

Consider, for instance, the equation set in Figure 2.1. The state $U_1 = 0$ permits two possible solutions for X and Y—$(X = 1, Y = 1)$ and $(X = 0, Y = 0)$—this equation set is not considered a causal model. Indeed, this equation set should not be a causal model, since it is under-defined. If U_1 has the value 0, then what should the value of X be? Should it rely on some kind of stochastic process? If so, we need to add another variable U_2 to model that process. Should this equation set model a type of flip-flop, such that X and Y have the value 1 if U_1 has the value 1, but retain their past values if U_1 has the value 0? If our intent is to describe a system that employs a notion of past states, then we need to define exactly what we mean by a "previous state" and formally define how such a previous state influences the present state. In Section 3.2, we offer such

a formalism, which we could use to precisely describe systems, such as flip-flops, that utilize the concepts of past history and previous state.

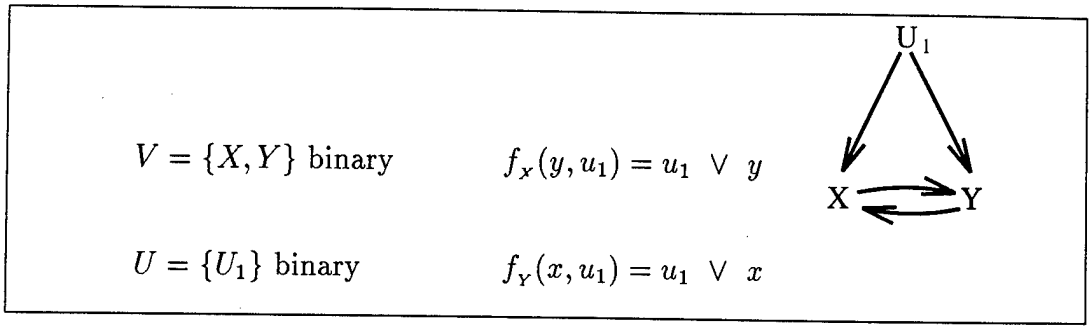


Figure 2.1: An equation set that is not a valid causal model

It is straightforward to show that every equation set that represents a recursive set of equations will necessarily be a causal model. The variables in a recursive equation set can be listed in order such that each variable is listed before any of its descendants. Since each variable is a deterministic function of its parents, each variable in the list is thus uniquely determined as long as all previous variables on the list are uniquely determined. By strong induction, since the first element in the list is uniquely determined by the variables U, all elements in the list must be uniquely determined by U.

Since only nonrecursive equation sets can be invalid causal models, it is useful to consider whether any nonrecursive equation set that is a valid causal model exists. There are, in fact, many nonrecursive equation sets that are also valid causal models. For example, consider the equation set in Figure 2.2.

This equation set is a valid causal model. The model, which denote as M dictates unique values for X and Y for both values of U_1: When $U_1 = 0$, X has the value 1 and Y has the value 0; $X(0) = 1$ and $Y(0) = 0$. When $U_1 = 1$, X

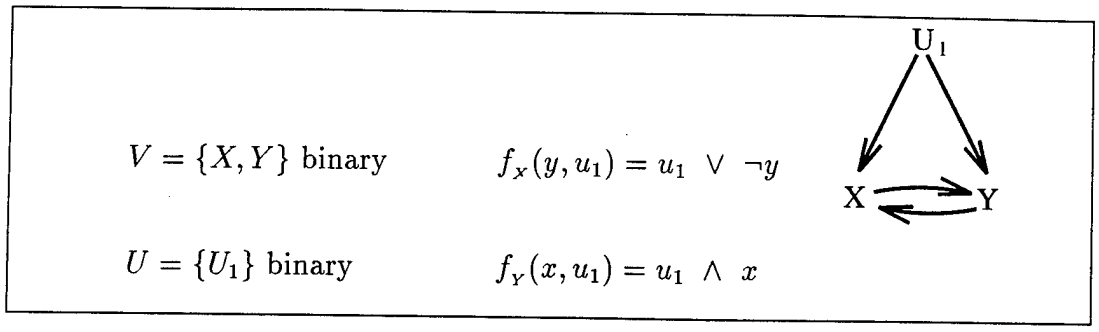


Figure 2.2: A valid non-recursive causal model

has the value 1 and Y has the value 1; $X(1) = 1$ and $Y(1) = 1$. In addition, the submodels of M also dictate unique solutions. There is a unique value for Y (for both values of U_1) in $M_{X=0}$ and $M_{X=1}$: $Y_{X=0}(0) = 0$, $Y_{X=0}(1) = 0$, $Y_{X=1}(0) = 0$, $Y_{X=1}(1) = 1$. There is also a unique value for X (for both values of U_1) in $M_{Y=0}$ and $M_{Y=1}$: $X_{Y=0}(0) = 1$, $X_{Y=0}(1) = 1$, $X_{Y=1}(0) = 0$, $X_{Y=1}(1) = 1$.

We can see that some nonrecursive equation sets are valid causal models. Does it make sense to say that these nonrecursive causal models actually model real-world phenomena? In some cases, yes, it does make sense. For instance, consider the common economic model of how supply and demand determine the price of a commodity. This is a nonrecursive causal model that models a real-world phenomena. The transient values are usually considered to be of no importance, so they are ignored. If the desired result is to model how the final price is obtained, a more complicated model is required. Such models are explored in section 3.3.

Finally, given that all linear equation sets are valid causal models, it is useful to consider just which nonlinear equation sets are also valid causal models. A nonrecursive equation set that is created by randomly selecting functions over binary variables will not often be a valid causal model. Such random equation

sets do not model any real-world phenomena, and as such are not of interest to us, and do not enter into our discussion of causal models.

The reason why we restrict every subset as well as the original set of the equations F to having a single fixed point will become clear as we describe interventions on causal models.

2.4 Interventions

Definition 2 merely provides a description of the mathematical objects that enter into a causal model. To fulfill our requirement that a causal model be capable of computing answers to causal queries, we need to supplement Definition 2 with an interpretation of the sentence "$X = x$ causes $Y = y$." In ordinary discourse, such a sentence implies that we can bring about the condition $Y = y$ by locally enforcing the condition $X = x$. Thus, Definition 2 must be supplemented with a formal interpretation of the notion "locally enforcing $X = x$" that is compatible with its common usage.

External intervention normally implies changing some mechanisms in the domain. In a logical circuit, for example, the act of enforcing the condition $X_i = 0$ by connecting some intermediate variable X_i to ground amounts to changing the mechanism that normally determines X_i. If X_i is the output of an OR gate, then after the intervention, X_i would no longer be determined by the OR gate but by a new mechanism (involving the ground) that clamps X_i to 0 regardless of the input to the OR gate. In the equational representation, this amounts to replacing the equation $x_i = f_i(\mathbf{pa}_i, u)$ with a new equation, $X_i = 0$, that represents the grounding of X_i.

The replacement of just one equation, not several, reflects the principle of locality in the common understanding of imperative sentences such as "Raise taxes" or "Make him laugh." When told to clean his face, a child does not ask for a razor, nor does he jump into the swimming pool. The proper interpretation of the modal sentence "*do p*" corresponds to a minimal perturbation of the existing state of affairs, and this, in the context of Definition 2, corresponds to the replacement of the minimal set of equations necessary to make p compatible with U.

In general, we will consider concurrent actions of the form $do(X = x)$, where X involves several variables in V.[1] This leads to the following definitions.

Definition 3 (submodel) *Let M be a causal model, X a set of variables in V, and x be a particular realization of X. A* submodel M_x *of M is the causal model*

$$M_x = \langle U, V, F_x \rangle$$

where

$$F_x = \{f_{X_i} : X_i \notin X\} \cup \{X = x\} \tag{2.2}$$

In words, F_x is formed by deleting from F all functions f_i corresponding to members of X and replacing these functions f_{X_i} with the set of functions $X = x$. Implicit in the definition of submodels is the assumption that F_x possesses a unique solution for every u.

Submodels are useful in representing the effects of local actions and changes. If we interpret each function f_i in F as an independent physical mechanism and define the action $do(X = x)$ as the minimal change in M required to make $X = x$ hold true under any u, then M_x represents the model that results from such a

[1] The formalization of conditional actions of the form "$do(X = x)$ if $Z = z$" is straightforward [Pea94].

minimal change, since it differs from M by only those mechanisms that determine the variables in X.

Definition 4 (effect of action) *Let M be a causal model, X be a set of variables in V, and x a particular realization of X. The* effect of action *$do(X = x)$ on M is given by the submodel M_x.*

Definition 5 (potential response) *Let Y be a variable in V, and let X be a subset of V. The* potential response *of Y to action $do(X = x)$, denoted $Y_x(u)$, is the solution for Y of the set of equations F_x.*

Definition 6 (counterfactual) *Let Y be a variable in V, and let X a subset of V. The* counterfactual *sentence "The value that Y would have obtained, had X been x" is interpreted as denoting the potential response $Y_x(u)$.*[2]

The syntactical transformation described in Definition 5 corresponds to replacing the old functional mechanisms $x_i = f_i(PA_i, u)$ with new mechanisms $X_i = x_i$ that represent the external forces that set the values x_i for each $X_i \in X$. As before, we assume each variable $Y \in V$ to be a unique function of the background U in any model M_x, namely, $Y = Y_{M_x}(u)$. For brevity, we will often omit the subscript M, leaving $Y_x(u)$.

An explicit translation of intervention into "wiping out" equations in the causal model was first proposed in [SW60] and used in [Fis70] and [Sob90]. Graphical ramifications are explicated in [SGS93] and [Pea93]. Interpretations of causal

[2]The connection between counterfactuals and local actions is made in [Lew73a] and is elaborated in [BP94] and [HS95]. Readers who are disturbed by the impracticality of actions in the interpretation of some counterfactuals (e.g., "If I were young") are invited to replace the word "action" with the word "modification" (see [Lea85]). [Pea95a, p. 706] explains the advantage of using hypothetical external interventions, rather than spontaneous changes, in thinking about causation and counterfactuals.

and counterfactual utterances in terms of $Y_x(u)$ are provided in [Pea96a]. Alternative formulations of causality, in terms of event trees, are given in [Rob86b] and [Sha96].

Note that $Y_x(u)$ is well defined even when $U = u$ and $X = x$ are incompatible in M (i.e., $X(u) \neq x$) Thus, there is room in the model for actions to enforce propositions that are not realized under normal conditions, or that are not realized under the abnormalities modeled in U. For example, if M describes a logic circuit we might wish to intervene and set some voltage X to x, even though the input dictates $X \neq x$. It is for this reason that one must invoke some notion of mechanism breakdown or "surgery" in the definition of interventions.

The unique feature of our formulation of actions—the feature that sets it apart from the formulations in control theory or decision analysis [Sav54, HS95]—is that an action is treated as a *modality*, namely, it is not given an explicit name but, rather, acquires the names of the propositions that it enforces as true. This enables the model to predict the effects of a huge number of action combinations without the modeler having to attend to such combinations. Instead, the causal model is constructed by specifying the characteristics of each individual mechanism under normal conditions, free of intervention.

2.5 Probabilistic Causal Models

If we wish to use causal models to do probabilistic reasoning, we need to add probability to the causal model framework. In probabilistic causal models, disturbances U are described with a probability distribution.

Definition 7 *(probabilistic causal model) A* probabilistic causal model *is a tuple*

$$\langle M, P(u) \rangle$$

Where

(i) M is a causal model $\langle V, U, F \rangle$

(ii) $P(u)$ is a probability distribution over U, such that each element $U_i \in U$ is marginally independent of all other elements of U.

The restriction on the probability distribution $P(u)$ that all members of U are independent does not limit the expressive power of probabilistic causal models. Any $P(u)$ with dependencies can be modeled by combining elements of U.

Given that each endogenous variable in a probabilistic causal model is a function of U and that a probabilistic causal model specifies a probability distribution over U, we can define a probability distribution over the endogenous variables in a probabilistic causal model. That is, for every set of variables $Y \subseteq V$, we have

$$P(y) = \sum_{\{u \mid Y(u)=y\}} P(u) \tag{2.3}$$

The probability of counterfactual statements is defined in the same manner, through the function $Y_x(u)$ induced by the submodel M_x:

$$P(Y_x = y) = \sum_{\{u \mid Y_x(u)=y\}} P(u) \tag{2.4}$$

We note that a causal model defines a joint distribution on all counterfactual statements, that is, $P(Y_x = y, Z_w = z)$ is defined for any sets of variables Y, X, Z, W, not necessarily disjoint. In particular, $P(Y_x = y, Y_{x'} = y')$ is well

defined and is given by $\sum_{u \mid Y_x(u)=y \ \& \ Y_{x'}(u)=y'} P(u)$. Likewise, $P(Y_x = y, X = x')$ is well defined and is given by $\sum_{u \mid Y_x(u)=y \ \& \ X(u)=x'} P(u)$.[3]

We can use probabilistic causal models to obtain a definition for the causal effect of one variable on another variable:

Definition 8 (causal effect) *Given two disjoint sets, $X \in V$ and $Y \in V$, the* causal effect *of X on Y is*

$$P(y|\hat{x}) = P(Y_x = y) \tag{2.5}$$

$$= \sum_{u \mid Y_x(u)=y} P(u) \tag{2.6}$$

for all values $x \in X$

In the statistical literature, causal effect is usually defined as the difference in expected values in Y, assuming $X = x$ or $X = x'$. Our definition is subtly different, since we consider the value for Y over the entire range of X, not just the difference between two values, x and x'.

2.6 Examples

Next we demonstrate the generality of the mathematical object defining causal models using two familiar applications: evidential reasoning and linear structural equation models.

2.6.1 Sprinkler Example

[3]The existence of such a joint distribution has prompted some of the objections to treating counterfactuals as random variables, because, when x and x' are incompatible, it is hard to attribute probability to the joint statement "Y would be y if X were x and X is actually x'." The definition of Y_x in terms of submodel not only avoids such problems but also illustrates that such joint probabilities can be encoded rather parsimoniously using $P(u)$ and F.

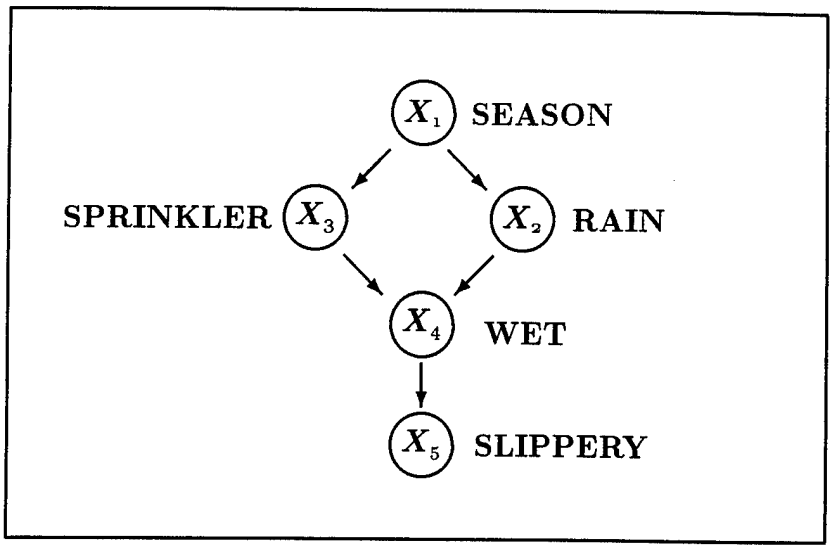


Figure 2.3: Causal graph illustrating causal relationships among five variables

Figure 2.3 is a simple yet typical causal graph used in commonsense reasoning. It describes the causal relationships among the season of the year (X_1), whether rain falls (X_2) during the season, whether the sprinkler is on (X_3), whether the pavement is wet (X_4), and whether the pavement is slippery (X_5). All variables in this graph except the root variable X_1 take a value of either "True" or "False." X_1 takes one of four values: "Spring," "Summer," "Fall," or "Winter." Here, the absence of a direct link between, for example, X_1 and X_5 captures our understanding that the influence of the season on the slipperiness of the pavement is mediated by other conditions (e.g., the wetness of the pavement). The corresponding model consists of five functions, each representing an autonomous mechanism:

$$\begin{aligned} x_1 &= u_1 \\ x_2 &= f_2(x_1, u_2) \\ x_3 &= f_3(x_1, u_3) \end{aligned}$$

$$
\begin{aligned}
x_4 &= f_4(x_3, x_2, u_4) \\
x_5 &= f_5(x_4, u_5)
\end{aligned} \tag{2.7}
$$

The disturbances $U_1, \ldots, U_5$ are not shown explicitly in Figure 2.3 but are understood to govern the uncertainties associated with the causal relationships. The causal graph coincides with the Bayesian network associated with $P(x_1, \ldots, x_5)$ whenever the disturbances are assumed to be independent, $U_i \perp\!\!\!\perp U \setminus U_i$.

A typical specification of the functions $\{f_1, \ldots, f_5\}$ and the disturbance terms is given by the Boolean model

$$
\begin{aligned}
x_2 &= [(X_1 = \text{Winter}) \vee (X_1 = \text{Fall}) \vee ab_2] \wedge \neg ab_2' \\
x_3 &= [(X_1 = \text{Summer}) \vee (X_1 = \text{Spring}) \vee ab_3] \wedge \neg ab_3' \\
x_4 &= (x_2 \vee x_3 \vee ab_4) \wedge \neg ab_4' \\
x_5 &= (x_4 \vee ab_5) \wedge \neg ab_5'
\end{aligned} \tag{2.8}
$$

where x_i stands for $X_i = true$, and ab_i and ab_i' stand, respectively, for triggering and inhibiting abnormalities. For example, ab_4 stands for (unspecified) events that might cause the pavement to get wet (x_4) when the sprinkler is off ($\neg x_2$) and it does not rain ($\neg x_3$) (e.g., pouring a pail of water on the pavement), while $\neg ab_4'$ stands for (unspecified) events that will keep the pavement dry ($\neg x_4$) in spite of rain falling (x_3), the sprinkler being on (x_2), and ab_4 (e.g., covering the pavement with a plastic sheet).

To represent the action "turning the sprinkler ON," or $do(X_3 = \text{ON})$, we replace the equation $x_3 = f_3(x_1, u_3)$ in the model of Eq. (2.7) with $X_3 = \text{ON}$. The resulting submodel, $M_{X_3=\text{ON}}$, contains all the information needed for computing the effect of the action on the other variables. It is easy to see from this submodel that the only variables affected by the action are X_4 and X_5, that is, the

descendants of the manipulated variable X_3. Note, however, that the operation $do(X_3 = \text{ON})$ stands in marked contrast to that of *finding* the sprinkler ON; the latter involves making the substitution $X_3 = \text{ON}$ *without* removing the equation for X_3, and therefore may potentially influence (the belief in) every variable in the network. This mirrors the difference between seeing and doing: after observing that the sprinkler is ON, we may wish to infer that the season is dry, that it probably did not rain, and so on; no such inferences can be drawn about the reasons for the action "turning the sprinkler ON."

2.6.2 Policy Analysis in Linear Econometric Models

Causal models are often used to predict the effect of policies on systems in dynamic equilibrium. In the economic literature, for example, we find the system of equations

$$q = b_1 p + d_1 i + u_1 \tag{2.9}$$

$$p = b_2 q + d_2 w + u_2 \tag{2.10}$$

where q is the quantity of household demand for a product A, p is the unit price of product A, i is household income, w is the wage rate for producing product A, and u_1 and u_2 represent error terms, namely, unmodeled factors that affect quantity and price, respectively [Gol92].

This system of equations constitutes a causal model (Definition 2) if we define $V = \{Q, P\}$, $U = \{U_1, U_2, I, W\}$ and assume that each equation represents an autonomous process in the sense of Definition 4. The causal graph of this model is shown in Figure 2.4. It is normally assumed that I and W are known, while U_1 and U_2 are unobservable and independent in I and W. Since the error terms

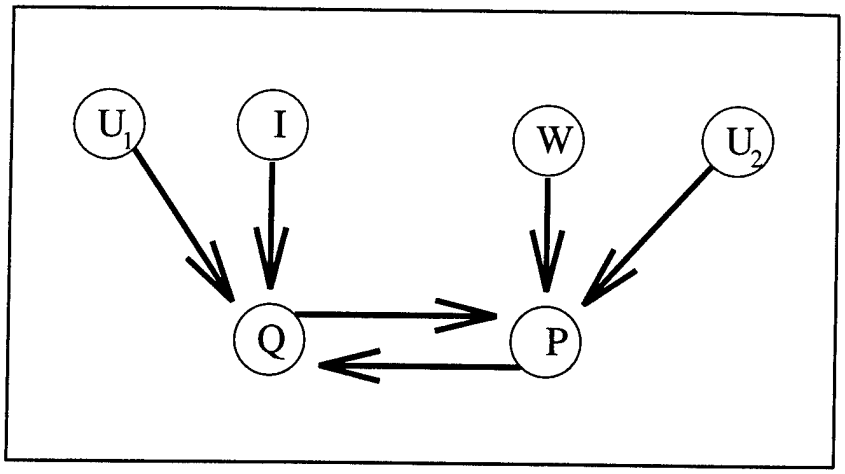


Figure 2.4: Causal graph illustrating the relationship between supply and demand

U_1 and U_2 are unobserved, the model must be augmented with the distribution of these errors, which is usually taken to be a Gaussian distribution with the covariance matrix $\Sigma_{ij} = cov(u_i, u_j)$.

We can use this model to answer queries such as:

1. Find the expected demand (Q) if the price is controlled at $P = p_0$.
2. Find the expected demand (Q) if the price is reported to be $P = p_0$.
3. Given that the current price is $P = p_0$, find the expected demand (Q) had the price been $P = p_1$.

To find the answer to the first query, we replace Eq. (2.10) with $p = p_0$, leaving

$$q = b_1 p + d_1 i + u_1 \tag{2.11}$$

$$p = p_0 \tag{2.12}$$

The demand is then $q = p_0 b_1 + d_1 i + u_1$, and the expected value of Q can be obtained from i and the expectation of U_1, giving $E[Q|\hat{p}_0] = E[Q] + b_1(p - E[P]) + d_1(i - E[I])$.

The answer to the second query is given by conditioning Eq. (2.9) on the current observation $\{P = p_0, I = i, W = w\}$ and taking the expectation,

$$E[Q|p_0, i, w] = b_i p_0 + d_1 i + E[U_1|p_0, i, w]. \tag{2.13}$$

The computation of $E[U_1|p_0, i, w]$ is a standard procedure once Σ_{ij} is given [Med69]. Note that, although U_1 was assumed independent of I and W, this independence no longer holds once $P = p_0$ is observed. Note also that Eqs. (2.9) and (2.10) both participate in the solution and that the observed value p_0 will affect the expected demand Q (through $E[U_1|p_0, i, w]$) even when $b_1 = 0$, which is not the case in query 1.

The third query requires the conditional expectation of the counterfactual quantity $Q_{p=p_1}$, given the current observations $\{P = p_0, I = i, W = w\}$, namely,

$$E[Q_{p=p_1}|p_0, i, w] = b_1 p_1 + d_1 i + E[U_1|p_0, i, w] \tag{2.14}$$

The expected value $E[U_1|p_0, i, w]$ is the same in the solutions to the second and third queries; the latter differs only in the term $b_1 p_1$. A general method for solving such counterfactual queries is described in [BP95].

2.6.3 Linguistic Notions of Causality

Structural models provide a precise language for defining intuitive causal concepts. In this section, we provide some brief examples, all relating to a given structural model M.

- "X is a cause of Y" if there exist two values x and x' of X and a value u of U such that $Y_x(u) \neq Y_{x'}(u)$.

- "X is a cause of Y in context $Z = z$" if there exist two values x and x' of X and a value u of U such that $Y_{xz} \neq Y_{x'z}(u)$.

- "X is a direct cause of Y" if there exist two values x and x' of X and a value u of U such that $Y_{xr}(u) \neq Y_{x'r}(u)$ where r is some realization of $V \backslash X$.

The direct causes of a variable in a model are determined in part by the granularity of the model in question. For example, consider the causal model in Figure 2.3. The sprinkler is a not a direct cause of slippery pavement. However, we could change the granularity of the model, by removing the variable "wet pavement," and having the rain and sprinkler affect "slippery pavement" directly. Then the sprinkler would be a direct cause of slippery pavement. Likewise, in Figure 2.3, the sprinkler is a direct cause of wet pavement. If we added a variable for "airborne water" such that it was true if either the sprinkler was on or it was raining, and such that the pavement be wet if "airborne water" was true, then the sprinkler would no longer be a direct cause of wet pavement. All of these causal models model the same phenomena, but the set of direct causes is different, depending upon the granularity of the model.

- "X is an indirect cause of Y", if X is a cause of Y, and X is not a direct cause of Y. As in the previous example, the granularity of the model M will determine which causes are direct and indirect.

- "X is causally irrelevant to Y, given fixed Z" if $\forall u, z, x, x'$ $Y_{xz}(u) = Y_{x'z}(u)$ in every submodel of M_z. Causal irrelevance is an important concept that will be thoroughly explored in Chapter 4.

- "Event $X = x$ may have caused $Y = x$" if

  (i) $X = x$ and $Y = y$ are true, and

  (ii) There exists a value u of U such that $X(u) = x$, $Y(u) = y, Y_x(u) = y$ and $Y_{x'}(u) \neq y$ for some $x' \neq x$.

- "The unobserved event $X = x$ is a likely cause of $Y = y$" if

  (i) $Y = y$ is true, and

  (ii) $P(Y_x = y, Y_{x'} \neq y | Y = y)$ is high for some $x' \neq x$

- "Event $Y = y$ occurred despite $X = x$" if

  (i) $X = x$ and $Y = y$ are true, and

  (ii) $P(Y_x = y)$ is low.

The preceding list demonstrates that, by varying the quantifiers of U and X, we have the flexibility to find appropriate formalizations for many nuances of causal expressions.

2.7 Properties of Counterfactual Statements

We now provide some definitions and some properties that are true for all causal models.

2.7.1 Definitions of Causal Properties

Property 1 (composition) *For any two singleton variables Y and W and any set of variables X in a causal model,*

$$W_x(u) = w \Longrightarrow Y_{xw}(u) = Y_x(u) \tag{2.15}$$

Composition states that, in any context $Z = z$, if we force a variable to a value that it would have had without our intervention, then the intervention will have no effect on other variables in the system.

Since composition allows for the removal of a subscript (i.e., reducing $Y_{xw}(u)$ to $Y_x(u)$), we need an interpretation for a variable with an empty set of subscripts which, naturally, we identify with the variable under no interventions.

Definition 9 (null action) $Y_\emptyset(u) \doteq Y(u)$.

Corollary 1 (consistency) *For any variables Y and W in a causal model,*

$$W(u) = w \Longrightarrow Y(u) = Y_w(u) \tag{2.16}$$

Corollary 1 follows directly from composition and null action. The implication in Eq. (2.16) was called *consistency* by [Rob87].[4]

Property 2 (effectiveness) *For all variables Y and set of variables X in a causal model,*

$$Y_{xy}(u) = y \tag{2.17}$$

Effectiveness specifies the effect of an intervention on the manipulated variable itself, namely, that if we force a variable Y to have the value y, then regardless of other enforcements $X = x$, Y will indeed take on the value y.

[4]This property and composition are tacitly used in economics [Man90] and statistics within the so-called Rubin's model [Rub74]. To the best of our knowledge, Robins was the first to state consistency formally and to use it to derive other properties of counterfactuals [Rob87]. Composition was brought to our attention by Jamie Robins (personal communication, February 1995). A weak version of composition is mentioned explicitly in [Hol86, p. 968].

Property 3 (reversibility) *For any two variables Y and W and any set of variables X in a causal model,*

$$(Y_{xw}(u) = y) \;\&\; (W_{xy}(u) = w) \Longrightarrow Y_x(u) = y \tag{2.18}$$

Reversibility reflects memoryless behavior — the state of the system, V, tracks the state of U, regardless of U's history. Given a context $X = x$ as in Eq. (2.18), if forcing W to a value w results in a value y for Y and, in turn, forcing Y to y indeed results in $W = w$, then W and Y will have the values w and y, respectively, without any intervention. This follows from the requirement that the equations in every context $X = x$ have a unique solution. Thus, if we assume a solution $W = w$ and obtain $Y = y$ and, in turn, assuming a solution $Y = y$ yields $W = w$, then $(W = w, Y = y)$ is indeed the solution to the equations.

A typical example of irreversibility is a system of two agents who adhere to a tit-for-tat strategy (e.g., the prisoners' dilemma). Such a system has two stable solutions, cooperation and defection, under the same external conditions U, and thus it does not satisfy the reversibility condition; forcing either one of the agents to cooperate results in the other agent's cooperation $(Y_w(u) = y, W_y(u) = w)$, yet this does not guarantee cooperation from the start $(Y(u) = y, W(u) = w)$. Irreversibility, in such systems, is a product of using a state description that is too coarse, one in which all of the factors that determine the ultimate state of the system are not included in U. In a tit-for-tat strategy, the state description should include factors such as the previous actions of the players, and reversibility is restored once the missing factors are included.

In recursive systems, reversibility follows directly from composition. This can easily be seen by noting that in a recursive system, either $Y_{xw}(u) = Y_x(u)$ or $W_{xy}(u) = W_x(u)$. Thus, reversibility reduces to $(Y_{xw}(u) = y) \;\&\; (W_x(u) = w) \Longrightarrow$

$Y_x(u) = y$, which is another form of composition, or to $(Y_x(u) = y)$ & $(W_{xy}(u) = w) \Longrightarrow Y_x(u) = y$, which is trivially true. In nonrecursive systems, reversibility is a property of causal loops. If forcing W to a value w results in a value y for Y, and forcing Y to the value y results in W achieving the value w, then W and Y will have the values w and y, respectively, without any intervention.

2.7.2 Soundness of Composition, Effectiveness, and Reversibility

Following standard logic, we will consider a property of causal relationships to be *sound* if that property holds in all structural models.

Theorem 1 *Composition is sound.*

Proof:

Since $Y_x(u)$ has a unique solution, forming M_x and substituting out all other variables will yield a unique solution for Y, regardless of the order of substitution. So we will form M_x and examine the structural equation for Y in M_x, $Y_x = f_Y(x, z, w, u)$, where Z stands for the rest of the parent set of Y. To solve for Z, we substitute out all variables except X, Y, and W. In other words, we substitute out all variables in M_x without substituting into X, W, and Y, and express Z as a function of x, w, and u. We then plug this solution into f_Y to get $Y_x = f_Y(x, w, Z(x, w, u), u)$, which we can write as $Y_x = f(x, w, u)$. At this point, we can solve for W by substituting out all variables in M_X other than X, which leaves $Y_x = f(x, W(u, x), u)$. We can now see that if $w = W_x(u)$, then $Y_x(u) = Y_{xw}(u)$. □

This proof is still valid in cases where $X = \emptyset$.

Theorem 2 *Effectiveness is sound.*

Proof:

This theorem follows from Definition 2, where $Y_x(u)$ is interpreted as the unique solution for Y of a set of equations under $X = x$. □

Theorem 3 *Reversibility is sound.*

Proof:

Reversibility follows from the assumption that the solution for V in every submodel is unique. Since $Y_x(u)$ has a unique solution, forming M_x and substituting out all other variables will yield a unique solution for Y, regardless of the order of substitution. So we will form M_x and examine the structural equation for Y in M_x, which in general might be a function of X, W, U, and additional variables: $Y_x = f_Y(x, w, z, u)$, where Z stands for parents of Y not contained in $X \cup W \cup U$. We now solve for Z by substituting out all variables except X, Y, and W. That is, we substitute out all variables in M_x, without substituting into X, W, and Y and express Z as a function of x, w, and u. We then plug this solution into f_Y to get $Y_x = f_Y(x, w, Z(x, w, u), u)$, which we can write as $Y_x = f(x, w, u)$. We now consider what would happen if we solved for Y in M_{xw}. Since we avoided substituting anything into W when we solved for Y in M_x, we will get the same result as before, namely, $Y_{xw} = f(x, w, u)$. In the same way, we can show that $W_x = g(x, y, u)$ and $W_{xy} = g(x, y, u)$. So, solving for $y = Y_x(u)$, $w = W_x(u)$ is the same as solving for $y = f(x, w, u)$ and $w = g(x, y, u)$, which is the same as solving for $y = Y_{xw}(u)$, $w = W_{xy}(u)$. Thus, any solution y to $y = Y_{xw}(u), w = W_{xy}(u)$ is also a solution to $y = Y_x(u)$. □

2.7.3 Independence of Effectiveness, Composition, and Reversibility

We now show that effectiveness, composition, and reversibility are independent.

Theorem 4 *Effectiveness, composition, and reversibility are independent.*

Proof:

In nonrecursive systems, the three properties of composition, effectiveness, and reversibility are independent—none is a consequence of the other two. This can be shown by constructing a truth table for counterfactual statements, such that any two properties hold and the third does not. Consider a model of two binary variables X and Y, and a single value u of U. Table 2.1 is a truth table for counterfactual statements over X and Y such that composition and effectiveness hold but reversibility does not. Table 2.2 is a similar table where effectiveness and reversibility hold but composition does not. Finally, table 2.3 is a truth table where composition and reversibility hold but effectiveness does not.

2.7.4 Completeness of Causal Properties

Examining these properties of causal models raises two obvious questions. One is "Have we missed any?" That is, are there any restrictions on legal causal sentences that are not captured by the above properties? The other is "How are these properties different from those derived in other systems?" That is, does our formalism impose more restrictions than other systems on the number of valid causal statements? How does the number of valid causal statements in our system compare, for instance, to the number in the very general one of Lewis? This section will answer both of these questions, at least for the case of recursive models.

$X=0$	$Y=0$		
$X_{X=0}=0$	$Y_{X=0}=0$	$X_{X=0,Y=0}=0$	$Y_{X=0,Y=0}=0$
$X_{X=1}=1$	$Y_{X=1}=1$	$X_{X=0,Y=1}=0$	$Y_{X=0,Y=1}=1$
$X_{Y=0}=0$	$Y_{Y=0}=0$	$X_{X=1,Y=0}=1$	$Y_{X=1,Y=0}=0$
$X_{Y=1}=1$	$Y_{Y=1}=1$	$X_{X=1,Y=1}=1$	$Y_{X=1,Y=1}=1$

Table 2.1: Table of counterfactual statements where composition and effectiveness hold but reversibility does not

$X=0$	$Y=1$		
$X_{X=0}=0$	$Y_{X=0}=1$	$X_{X=0,Y=0}=0$	$Y_{X=0,Y=0}=0$
$X_{X=1}=1$	$Y_{X=1}=0$	$X_{X=0,Y=1}=0$	$Y_{X=0,Y=1}=1$
$X_{Y=0}=0$	$Y_{Y=0}=0$	$X_{X=1,Y=0}=1$	$Y_{X=1,Y=0}=0$
$X_{Y=1}=1$	$Y_{Y=1}=1$	$X_{X=1,Y=1}=1$	$Y_{X=1,Y=1}=1$

Table 2.2: Table of counterfactual statements where effectiveness and reversibility hold but composition does not

$X = 0$	$Y = 1$		
$X_{X=0} = 0$	$Y_{X=0} = 1$	$X_{X=0,Y=0} = 0$	$Y_{X=0,Y=0} = 1$
$X_{X=1} = 0$	$Y_{X=1} = 1$	$X_{X=0,Y=1} = 0$	$Y_{X=0,Y=1} = 1$
$X_{Y=0} = 0$	$Y_{Y=0} = 1$	$X_{X=1,Y=0} = 0$	$Y_{X=1,Y=0} = 1$
$X_{Y=1} = 0$	$Y_{Y=1} = 1$	$X_{X=1,Y=1} = 0$	$Y_{X=1,Y=1} = 1$

Table 2.3: Table of counterfactual statements where composition and reversibility hold but effectiveness does not

Lewis was very careful to keep his formalism as general as possible, and, save for the obvious requirement that every world be closest to itself, he did not impose any specific structure on the distance measure. However, the fact that people manage to communicate with counterfactuals suggests that the distance measure is shared by many people and, hence, that it is not entirely arbitrary but must be one which can be encoded parsimoniously in the mind. What then is the mental representation used in the encoding of interworld distances?

Lewis himself provides a clue: the closest worlds that he envisions are *causal* in nature. For instance, when Lewis considers as an example a hypothetical world in which kangaroos have no tails, he argues that not just the state of the tail, but also the tracks that the animal makes, the animal's balance, and a variety of other factors would also be different. Thus, Lewis appeals to our common knowledge of cause and effect in laying out which factors are expected to be different in the

hypothetical world and which factors are expected to be unaltered.

If our assessment of interworld distances comes from causal knowledge, the question of whether that knowledge imposes its own structure on distances, a structure that is not captured in Lewis's logic, arises. Phrased differently, by agreeing to measure closest worlds on the basis of causal relations, do we restrict the set of counterfactual statements we regard as valid. The question is not merely theoretical. For example, Gibbard and Harper [1981] characterize sentences of the form "If we do A, then B"using Lewis's general framework, while Pearl [Pea95a] developed a calculus of action based directly on causal models— are the two formalisms are identical?

For recursive systems, the answer is yes. As we prove next, given a causal ordering on the variables in the system, composition is complete.[5] Thus, for recursive systems, we know that we have not missed any causal properties, and that our formalism imposes no more restrictions than Lewis's on the validity of causal or counterfactual statements.

Theorem 5 *Composition, together with effectiveness, definiteness, and uniqueness, are complete for causally ordered systems.*

We first give some notational definitions that we will use in our proof. A formal proof of completeness requires two additional properties, definiteness and uniqueness.[6]

[5]The formal completeness proof requires effectiveness and two other technical definitions. Composition (and, for nonrecursive systems, reversibility) encapsulate the essence of structural models.

[6]These two properties, definiteness and uniqueness, were kept implicit in the completeness proof originally reported in [GP97]; the need to explicate them formally was brought to my attention by [Hal97].

Property 4 (definiteness) *For any variable X and set of variables Y,*

$$\exists x \in X \ s.t. \ \ X_y(u) = x \tag{2.19}$$

Property 5 (uniqueness) *For every variable X and set of of variables Y,*

$$X_y(u) = x \ \& \ X_y(u) = x' \Longrightarrow x = x' \tag{2.20}$$

Definition 10 (statement) *By a counterfactual statement, or* statement *for short, we denote a sentence of the form $Y_x(u) = y$ for a specific variable $Y \in V$, a specific realization x of $X \subseteq V$, and a specific u in the domain of U.*

Definition 11 (causal ordering) *A* causal ordering $X_1 \ldots X_n$ *of a set of variables is an ordering such that for any two variables $X = X_i$ and $Y = X_k$ such that $i < k$, $X_{yz}(u) = X_z(u)$, where Z is any set of variables not including X or Y.*

Clearly, for every recursive model we can find an ordering that satisfies the condition of Definition 11. In fact, every ordering consistent with the arrows of the causal graph $G(M)$ will satisfy this condition. A system in which the variables are indexed along a specific causal ordering will be called a *causally ordered system.*

Definition 12 (semantic entailment) *Given a set S of counterfactual statements, let M_S be the set of models of S, namely, the set $\{m_1, \ldots, m_n\}$ of all causal models such that all statements in S hold for each m_i. A counterfactual statement σ is* semantically entailed *by S, written $S \models \sigma$, if σ holds in each $m_i \in M_S$.*

Definition 13 (syntactic entailment) *Given a set A of axioms, a set of counterfactual statements S* syntactically entails *a counterfactual statement σ, written*

$S \vdash_A \sigma$, if σ can be derived from S using repeated applications of axioms from A together with the rules of logic.

Define A_C to be the set {composition, effectiveness, definiteness, uniqueness, causal ordering}. We want to show that all statements that are semantically entailed by S are also syntactically entailed by S, namely, that

$$S \models \sigma \implies S \vdash_{A_C} \sigma$$

It is enough to show that every set of statements S that is consistent with A_C has a model. To see that this is sufficient to prove the completeness of A_C, assume that there is some set S and statement $p : X_z(u) = x$ such that in every model consistent with S, p holds, and p is not derivable from S using A_C. Since p is not derivable from S, there must be some other statement $p' : X_z(u) = x', x \neq x'$, such that $S \cup \{p'\}$ is consistent with A_C. Since in every model consistent with S, $X_z(u) = x$ holds, no model is consistent with $S \cup \{p'\}$. Thus, if A_C is not complete, then there must exist some set S' that is consistent with A_C, and has no model. Looking at the contrapositive, if every set of statements S that is consistent with A_C has a model, then A_C is complete.

We now show that for any set of statements S, if S is consistent under A_C then S has a model. We will use the concept of a maximally consistent set, which is a standard technique used to prove completeness in modal logic [FHM95]. Consider a maximally consistent set S^*. That is, a superset of S that is consistent with A_C such that any superset of S^* is not consistent with A_C. We will show that there is a causal model M which satisfies every statement in S^*, and thus satisfies every statement in S.[7]

[7]We thank Joseph Halpern for calling my attention to this technique which simplifies appreciably the completeness proof originally reported in [GP97].

Proof (by induction): We prove that, for any maximally consistent set S^*, there exists a causal model M which satisfies every statement in S^*, by induction on the number of variables $|V|$ in S^*.

Base Case:

If $|V| = 1$, then the statements $X(u)$ in S^* determine the function for X, and effectiveness ensures that $X_x(u) = x$ for all $x \in X$.

Inductive Case:

Consider the variables V that are in S^*. Let $Y \in V$ be the last element in the causal ordering. Consider the set S'^*, which is S^* with all statements of the form $Y_z(u) = y$ and $X_{yz}(u) = x$ removed. By the inductive hypothesis, there is a model M' such that every element of S'^* is satisfied.

We now extend M' to M, such that every element in S^* is satisfied in M. For each variable $X \in M'$ and each value y of Y, $f_{XM}(x_1, \ldots, x_k, y, u) = f_{XM'}(x_1, \ldots, x_k, u)$. We define f_Y as follows: for each statement $(Y_z(u) = y) \in S^*$ such that $|Z| = |V| - 1$ and $Y \notin Z$, $f_Y(z, u) = y$. Definiteness ensures that f_Y will be completely determined.

Since M' satisfied all elements of S'^*, and given the causal ordering such that $X_{yz}(u) = X_z(u)$ for all $X_{yz}(u), X_z(u)$ in S^*, M satisfies all statements of the form form $X_z(u)$ in S^*.

We now show that M satisfies every element of S^* of the form $Y_z(u) = y$. We show this by induction on the size of $|V| - |Z|$.

Base Cases:

(i) $Y \in Z$. By effectiveness, $Y_z(u) = y$ is in M.

(i) $|V| - |Z| = 1$. By construction of f_Y, $Y_z(u) = y \Longrightarrow Y = y$ is in M_z.

Inductive Case:

$|V| - |Z| = k$. Consider $Y_{zx}(u) = y'$, where $x = X_z(u)$. Above, we proved that $X_z(u)$ is satisfied in M, and by the inductive hypothesis, $Y_{zx}(u) = y'$ is satisfied in M. Thus, by composition, $Y_z(u) = y'$ is satisfied in M and, also by composition, $y = y'$. Thus, $Y_z(u) = y$ is satisfied in M. □

Joseph Halpern [Hal97] has recently shown that composition, reversibility, effectiveness, and definiteness are complete in all causal models, recursive as well as nonrecursive, as long as the uniqueness assumption holds. He further characterized systems in which uniqueness does not hold, which using a more elaborate type of axioms.

2.8 Comparison to Lewis's Formalism

We now compare our causal model framework to that of Lewis [Lew73b], to show that for recursive systems, composition and effectiveness are sound and complete within Lewis's framework. We give here a version of Lewis's logic for counterfactual sentences (from [Lew81]).

Rules

(1) If A and $A \Longrightarrow B$ are theorems, so is B

(2) If $(B_1 \,\&\, \ldots) \Longrightarrow C$ is a theorem, then so is

$$((A \,\Box\!\!\rightarrow B_1)\ldots) \Longrightarrow (A \,\Box\!\!\rightarrow C)$$

Axioms

(1) All truth-functional tautologies

(2) $A \,\Box\!\!\rightarrow A$

(3) $(A \,\Box\!\!\rightarrow B)\,\&\,(B \,\Box\!\!\rightarrow A) \Longrightarrow (A \,\Box\!\!\rightarrow C) \equiv (B \,\Box\!\!\rightarrow C)$

(4) $((A \vee B) \,\Box\!\!\rightarrow A) \vee ((A \vee B) \,\Box\!\!\rightarrow B) \vee (((A \vee B) \,\Box\!\!\rightarrow C) \equiv (A \,\Box\!\!\rightarrow C)$

$\&\ (B \,\Box\!\!\rightarrow C))$

(5) $A \,\Box\!\!\rightarrow B \Longrightarrow A \Longrightarrow B$

(6) $A \,\&\, B \Longrightarrow A \,\Box\!\!\rightarrow B$

The statement $A \,\Box\!\!\rightarrow B$ stands for "In all closest worlds where A holds, B holds as well." Lewis does not put any restrictions on the definition of closest worlds, beyond the obvious requirement that world w be no further from itself than any other world $w' \neq w$. In essence, causal models with local interventions define an ordering among worlds that gives a metric by which to define what worlds are

closest. As such, all of Lewis's axioms are true for causal models and follow from effectiveness, composition, and (for nonrecursive systems) reversibility.

In order to relate Lewis's axioms to our framework, we need to translate his syntax into the language of causal models. We will equate Lewis's "world" with an instantiation of all variables in a causal model, including the variables in U. Propositions, such as A and B in the statements above, will be limited to the assignment of values to subsets of variables in a model. Thus, the meaning of the statement $A \Box\!\!\rightarrow B$ in causal models is "If we force a set of variables to have the values A, a second set of variables will have the values B." Let A stand for a set of values $x_1, \ldots, x_n$ of the variables $X_1, \ldots, X_n$, and let B stand for a set of values $y_1, \ldots, y_m$ of the variables $Y_1, \ldots, Y_m$. Then,

$$\begin{aligned} A \Box\!\!\rightarrow B \quad \equiv \quad & Y_{1x_1 \ldots x_n}(u) = y_1 \ \& \\ & Y_{2x_1 \ldots x_n}(u) = y_2 \ \& \\ & \ldots \\ & Y_{2x_1 \ldots x_n}(u) = y_n \ \& \end{aligned} \tag{2.21}$$

Conversely, we need to define what statements such as $Y_x(u) = y$ mean in Lewis's notation. Let A stand for the proposition $X = x$, and B stand for the proposition $Y = y$. Then,

$$Y_x(u) = y \equiv A \Box\!\!\rightarrow B \tag{2.22}$$

We can now examine each of Lewis's axioms in turn.

(1) Trivially true.

(2) This axiom is the same as effectiveness. Namely, if we force a set of variables X to have the value x, then the resulting value of X is x. That is, $X_x(u) = x$.

(3) This axiom is a weaker form of reversibility, which is relevant only for nonrecursive causal models.

(4) Since actions in causal models are restricted to conjunctions of literals, this axiom does not apply. However, under the interpretation $do(A \vee B) \equiv do(A) \vee do(B)$, this axiom does hold.

(5) This axiom follows directly from composition.

(6) This axiom follows directly from composition.

Likewise, composition and effectiveness follow from Lewis's axioms. Composition is a consequence of Lewis's axiom (5) and rule (1), while effectiveness is the same as Lewis's axiom (2). Thus, causal models do not add any restrictions to counterfactual statements above those imposed by Lewis's framework, when we are considering recursive models. When we consider nonrecursive systems, we see that reversibility is not enforced by Lewis's framework. Lewis's axiom (3), while similar, is not as strong as reversibility. For instance, $Y = y$ may hold in all closest w-worlds, $W = w$ may hold in all closest y-worlds and, still, $Y = y$ may not hold in our world. A graphical example violating reversibility in Lewis' framework is given in Figure 2.5

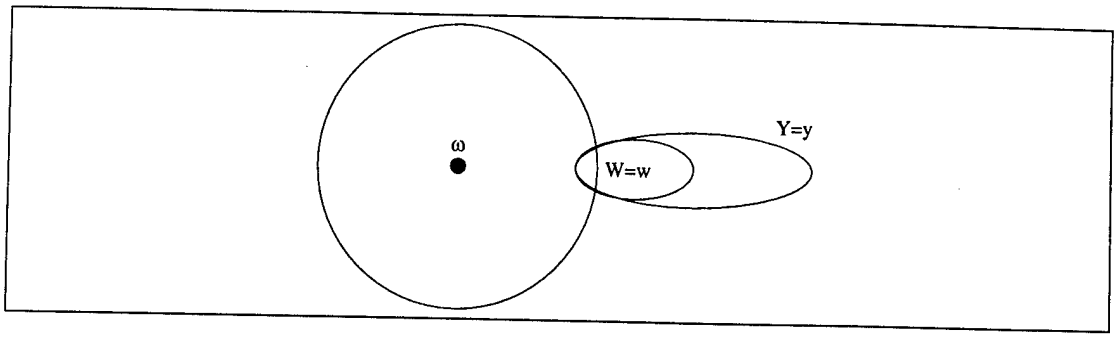


Figure 2.5: Example of the failure of reversibility in Lewis's framework: $W = w$ holds in all closest y-worlds, and $Y = y$ holds in all closest w-worlds, yet $Y \neq Y$ and $W \neq w$

2.9 Applying Counterfactual Derivation: Example

Consider the century-old debate about the effect of smoking on the incidence of lung cancer. According to many, the tobacco industry has managed to block anti-smoking legislation by arguing that the observed correlation between smoking (X) and lung cancer (Y) could be explained by some sort of carcinogenic genotype (U_1) that involves inborn craving for nicotine.[8] However, according to the Surgeon General's report of 1964, there is a causal link between smoking and lung cancer that is mediated by the accumulation of tar deposits in a person's lungs (Z). The two claims are combined in the graphical model of Figure 2.6, which represents causal theories having the following structure:

$$V = \{X \text{ (Smoking)}, Y \text{ (Lung Cancer)}, Z \text{ (Tar in Lungs)}\}$$

$$U = \{U_1, U_2\} \quad (U_1 \perp\!\!\!\perp U_2)_{P(u)}$$

$$X = f_1(U_1), \; Z = f_2(X, U_2), \; Y = f_3(Z, U_1)$$

The graphical model embodies several assumptions. The absence of a direct

[8] For an excellent historical account of this debate, see [SGS93, pp. 291–302].

link between X and Y represents the assumption that the effect of smoking cigarettes (X) on the production of lung cancer (Y) is entirely mediated through tar deposits in the lungs (Z). The lack of a direct link between U_1 and U_2 reflects the assumption that even if a genotype is aggravating the production of lung cancer, it nevertheless has no effect on the amount of tar in the lungs except indirectly, through cigarette smoking.

The graph conveys in fact the stronger assumption that U_1 and U_2 are marginally independent , written $(U_1 \perp\!\!\!\perp U_2)_{P(u)}$, which is represented by the absence of a dotted arc connecting U_1 and U_2.

To demonstrate how we can assess the degree to which cigarette smoking increases (or decreases) lung cancer risk, we imagine a study in which the three variables, X, Y, and Z, were measured simultaneously on a large, randomly selected sample from the population. From such data, we wish to assess the risk of lung cancer (for a randomly chosen person in the population) under two hypothetical policies: smoking ($X = 1$) and refraining from smoking ($X = 0$). In other words, we wish to derive an expression for the probability of the causal effect Y_x , $P(Y_x = y)$, based on the joint distribution $P(x, y, z)$ and the assumptions embedded in the graphical model. These assumptions can be translated into the language of counterfactuals, using two simple rules (see [Pea95a, p. 704]):

Rule 1 *Exclusion restrictions.* For every variable Y having parents PA_Y, and for every set of variables Z disjoint of PA_i, we have

$$Y_{pa_Y}(u) = Y_{pa_Y z}(u) \tag{2.23}$$

Rule 2 *Independence restrictions.* If $Z_1, \ldots, Z_k$ is any set of nodes in V not con-

nected to Y via some U variable, we have

$$Y_{pa_Y} \perp\!\!\!\perp \{Z_{1pa_{Z_1}}, \ldots, Z_{kpa_{Z_k}} \tag{2.24}$$

Terms not parameterized by u, such as those in Eq. (2.24), denote random variables induced by $P(u)$.

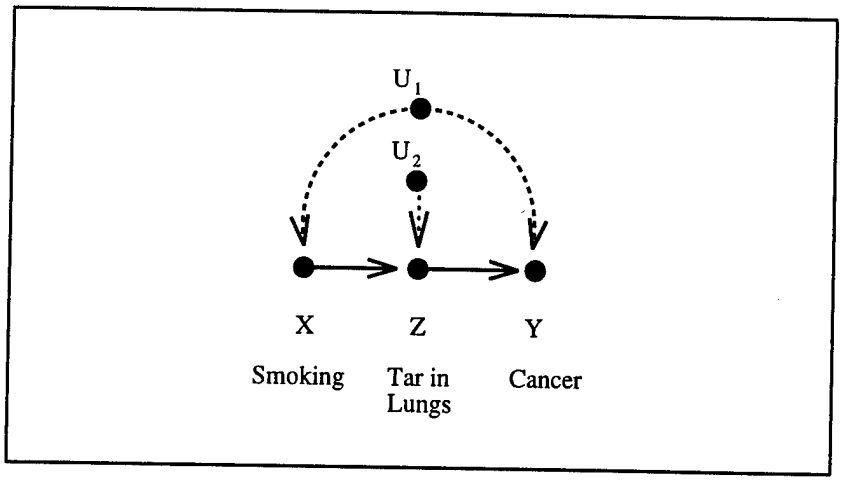


Figure 2.6: Causal graph illustrating the effect of smoking on lung cancer

Applying these two rules, we see that the causal model encodes the following assumptions:

$$Z_x(u) = Z_{yz}(u) \tag{2.25}$$

$$X_y(u) = X_{zy}(u) = X_z(u) = X(u) \tag{2.26}$$

$$Y_z(u) = Y_{zx}(u) \tag{2.27}$$

$$Z_x \perp\!\!\!\perp \{Y_z, X\} \tag{2.28}$$

Eqs. (2.25)–(2.27) were obtained using the exclusion restrictions of Eq. (2.23). Eq. (2.25), for instance, represents the absence of a directed path from Y to X, while Eq. (2.26) represents the absence of a causal link from Z or Y to X. In

contrast, Eq. (2.28) follows from the independence restriction of Eq. (2.24) and represents the lack of a connection between (i.e., the independence of) U_1 and U_2.

We now use these assumptions, and the properties of composition and effectiveness, to compute various tasks:

Task 1 Compute $P(Z_x = z)$, the probabilistic causal effect of X on Z.

$$\begin{aligned} P(Z_x = z) &= P(Z_x = z|x) \quad \text{from Eq. (2.28)} \\ &= P(Z = z|x) \quad \text{by composition} \\ &= P(z|x) \end{aligned} \tag{2.29}$$

Task 2 Compute $P(Y_z = y)$, the probabilistic causal effect of Z on Y.

$$P(Y_z = y) = \sum_x P(Y_z = y|x)P(x) \tag{2.30}$$

and since $Y_z \perp\!\!\!\perp Z_x|X$,

$$\begin{aligned} P(Y_z = y|x) &= P(Y_z = y|x, Z_x = z) \quad \text{from Eq. (2.28)} \\ &= P(Y_z = z|x, z) \quad \text{by composition} \\ &= P(y|x, z) \quad \text{by composition} \end{aligned} \tag{2.31}$$

Substituting Eq. (2.31) in Eq. (2.30) gives

$$P(Y_z = y) = \sum_x P(y|x, z)P(x) \tag{2.32}$$

Task 3 Compute $P(Y_x = y)$, the probabilistic causal effect of X on Y.

For any variable Z,

$$Y_x(u) = Y_{xz}(u), \quad \text{if} \quad Z_x(u) = z \qquad \text{by composition}$$

Since $Y_{xz}(u) = Y_z(u)$ (from Eq. (2.33)),

$$Y_x(u) = Y_{xz_x}(u) = Y_z(u) \quad \text{when} \quad z_x = Z_x(u) \tag{2.33}$$

Thus,

$$\begin{aligned} P(Y_x = y) &= P(Y_{z_x} = y) && \text{from Eq. (2.33)} \\ &= \textstyle\sum_z P(Y_{z_x} = y | Z_x = z) P(Z_x = z) \\ &= \textstyle\sum_z P(Y_z = y | Z_x = z) P(Z_x = z) && \text{by composition} \\ &= \textstyle\sum_z P(Y_z = y) P(Z_x = z) && \text{from Eq. (2.28)} \end{aligned} \tag{2.34}$$

$P(Y_x = y)$ and $P(Z_x = z)$ were computed in Eq. (2.29) and Eq. (2.32). Substituting gives us

$$P(Y_x = y) = \sum_z P(z|x) \sum_{x'} P(y|z, x') P(x') \tag{2.35}$$

The right hand side of Eq. (2.35) can be computed from $P(x, y, z)$.

In general, the reduction of counterfactual probabilities $P(Y_x = y)$ to expressions involving probabilities over observed variables is called *identifiability*. Our completeness result implies that any identifiable counterfactual quantity $P(Y_x = y)$ can be reduced to the correct expression by repeated application of composition and effectiveness.

2.10 Conclusion

We have given a formal, mathematical description of causal models, as well as some justification for our model. We have derived a set of properties of causation—composition, effectiveness, and reversibility—that follow from the structural model formalism. For recursive (and causally ordered) models, we have shown that composition and effectiveness are complete.

The completeness proof for composition and effectiveness in recursive causal models has two major implications for recursive systems. First, it shows that the structural interpretation of counterfactuals adds no restrictions beyond those of Lewis's closest-world interpretation. Second, it shows that the very general language of Lewis's closest-world framework embodies all of the causal restrictions on counterfactuals that are not embodied already by the requirement of recursiveness. In nonrecursive systems, however, there is a difference between the two formalisms. The causal reading of counterfactuals imposes the additional restriction of reversibility.

Moreover, the completeness result assures us that a proof may safely be attempted with two axioms only, that is, all truths derivable by graph-based analysis are also derivable using effectiveness and composition. This does not in any way diminish the usefulness of graphs in causal analysis. Graphs play an essential role in knowledge specification—graphical specification of premises followed by translation to counterfactuals is more natural than trying to articulate premises directly as counterfactuals. Graphs may also assist in the proof procedure and in providing independence relations (among counterfactuals and visible variables) that are not easily derived symbolically. Nevertheless, we have shown that the power of symbolic counterfactual analysis is not lower than that of graphical

analysis whenever the system is causally ordered. In nonrecursive models, this is not necessarily the case. Attempts to evaluate counterfactual statements using only composition and effectiveness may fail to certify some statements that are true in all causal models, and whose validity can only be recognized through the use of reversibility. Whether composition, effectiveness, reversibility are complete for nonrecursive as well as recursive systems has remained an open question until very recently. Halpern [Hal97] has settled the problem in the affirmative.

We have also shown how some standard causal utterances can be transformed into the structural model framework. This gives each of these terms a precise definition which, in turn, and allows us to reason about them without ambiguity.

CHAPTER 3

Dynamic Causal Models

3.1 Introduction

One problem with causal models is that they do not allow for the concept of memory. That is, each variable is a function of only the *current* state of other variables in the system, and, thus, only of the current value of the disturbances. The past cannot affect the present. Many real-world examples require the concept of previous state to be accurately modeled. For instance, a standard computer (which has an internal memory) could not be accurately modeled by a standard causal model. Even less complicated systems require the concept of past state. Something as simple as a ballpoint pen that whose point extends and retracts with a the push of single button is difficult to model with a standard causal model. Any system that requires the concept of a previous state will not be completely definable using a causal model. We can overcome this shortcoming by adding some simple extensions to causal models.

3.2 Causal Models with Memory

We will consider each variable X_i to be a function not only of the other V's and U's, but also of the previous state of a subset of variables in V. This subset may include the previous value of X_i as well as the previous values of the parents of X_i.

Our object is to define a smooth transition between full-blown temporal networks and static causal models. Often, temporal networks offer more power than we need. For instance, in economic models of supply and demand, we are interested only in the stable state of the variables, not in transitory effects. By removing these transitory effects from our model, we allow for simpler computations. Thus, we will improve the expressive power of causal models while maintaining the computational savings provided by the approximations of static systems.

More formally, we can define an extension to causal models as follows.

Definition 14 (causal model with memory) *A* causal model with memory *is a* 4*-tuple*

$$M = \langle V, V^-, U, F \rangle$$

where

(i) $V = \{X_1, \ldots, X_n\}$ *is a set of endogenous variables determined within the system,*

(ii) $V^- = \{X_1{}^-, \ldots, X_n{}^-\}$ *is a set of previous values of the endogenous variables* V*,*

(iii) $U = \{U_1, \ldots, U_m\}$ *is a set of exogenous variables that represent disturbances, abnormalities, assumptions, or boundary conditions, and*

(iv) $F = \{f_{X_i}\}$ *is a set of* n *deterministic, nontrivial functions, each of the form*

$$X_i = f_{X_i}(\mathbf{pa}_i, \mathbf{pa}^-{}_i, u) \quad i = 1, \ldots, n \quad \mathbf{pa}_i \subseteq V \setminus X_i, \mathbf{pa}^{\mathbf{i}}{}_i \subseteq V^-$$

We will assume that the set of equations in (iv) has a unique solution for $X_1, \ldots, X_n$*, given any value of the disturbances* $U_1, \ldots, U_m$ *and past values* V^-*.*

Thus we can consider each variable Y to be a function of the disturbances U and the past values of V in the causal model M: $Y = Y_M(u, v^-)$.

As with standard causal models, we can also define probabilistic causal models with memory.

Definition 15 (probabilistic causal model with memory) *A* probabilistic causal model with memory *is a tuple*

$$\langle M, P(u) \rangle$$

where

(i) M is a causal model with memory $\langle V, V^-, U, F \rangle$

(ii) $P(u)$ is a probability distribution over U, such that each element $U_i \in U$ is marginally independent of all other elements of U.

In this model, time is discretized. At each time step, the state of the world is determined by the previous state of each variable together with the current state of the other variables in the system.

Consider a ballpoint pen that extends and retracts its point with a single button. This item would be difficult, if not impossible, to model using a standard causal model. When the button is depressed and released, the current state of the point relies not on its current causal influences, but on its own previous state. Using a causal model with memory, however, the pen can easily be described, as in Figure 3.1.

Notice that in this example, $\mathbf{pa}^-_i = \{X_i^-\}$. We next consider what power we lose by restricting $\mathbf{pa}^-_i$ to X_i. Is there any system that cannot be modeled with this new restriction?

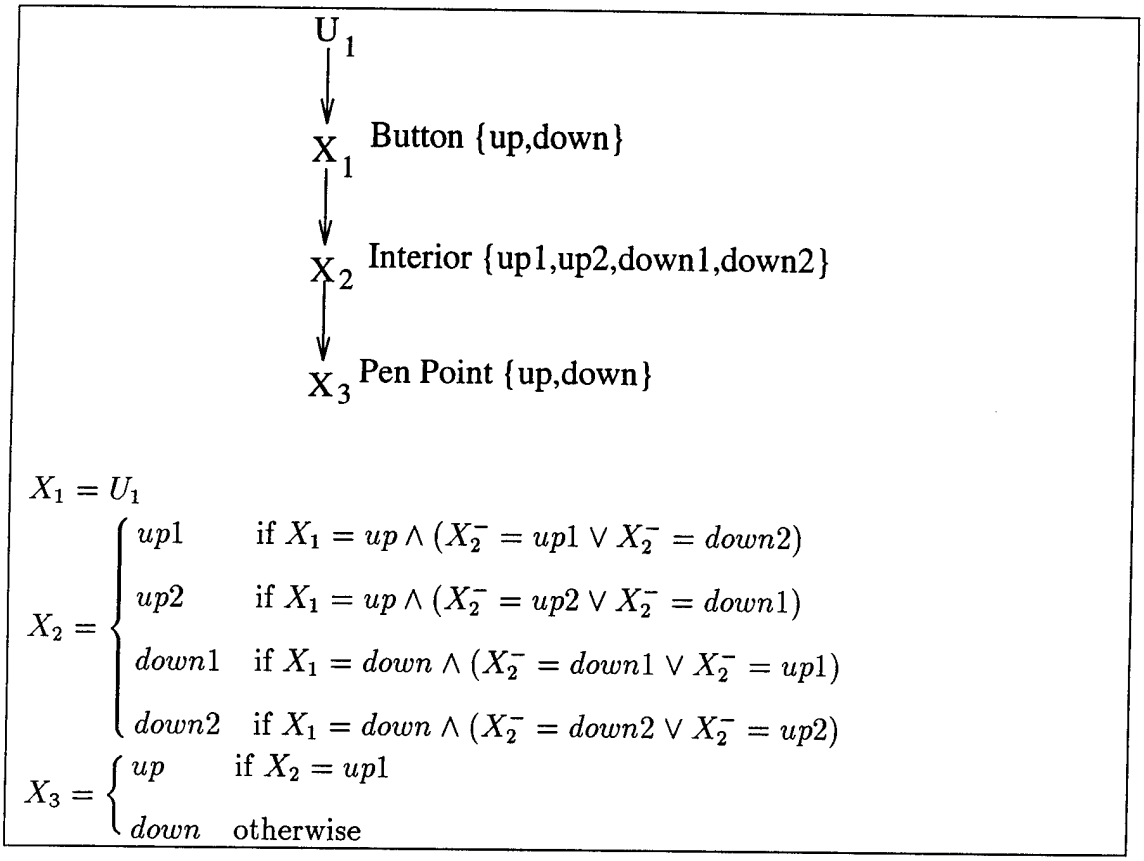


Figure 3.1: Example of a causal model with memory

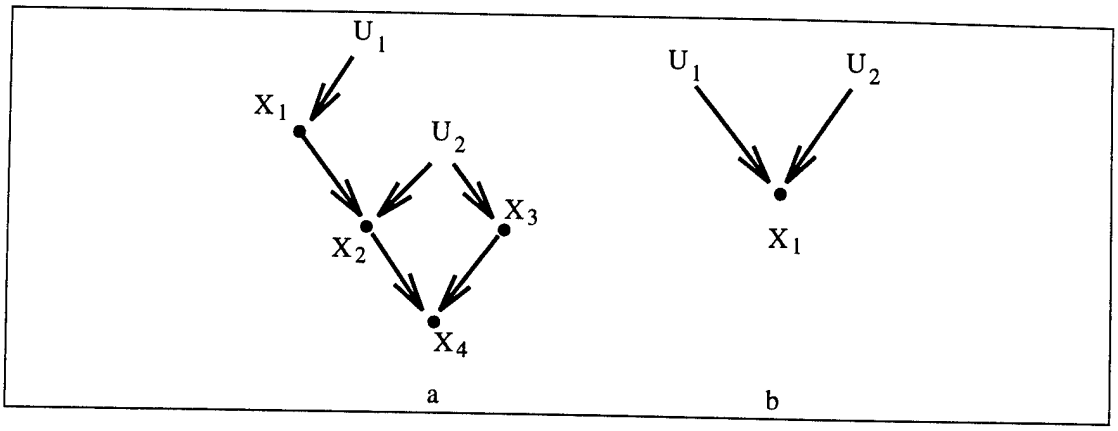


Figure 3.2: We can ensure that a $\mathbf{pa}^-_i = X_i$ in a causal model with memory by combining variables, in the extreme case, to a single variable.

If we do not limit the size of the domain of the variable involved, we can easily transform any causal model with memory into an equivalent model such that $\mathbf{pa}^-_i = X_i$ for all i. This can be done by combining variables into "meta"-variables, each of which has a larger domain. For instance, if a variable X_4 has X_2^- in $\mathbf{pa}^-_4$, we could combine X_4 and X_2 into one variable. In the extreme case, all of the variables in V can be combined into a single variable. An example of such a combination is given in Figure 3.2. If the variables $X_1, \ldots, X_4$ in Figure3.2a are all binary, the variables in the model can be combined to obtain Figure 3.2b, in which the single variable X_1 that has $2^4 = 16$ values.

This method of obtaining a causal model with memory such that $\mathbf{pa}^-_i = X_i$ suffers from two major disadvantages. One disadvantage is that the domains of the variables grow exponentially. More important, by combining variables in this way, we lose valuable information about the way the system behaves. We can no longer access information on how the variables in the system respond to intervention, which is precisely why we developed causal models in the first place.

We can, however, solve the problem without losing the basic structure of the

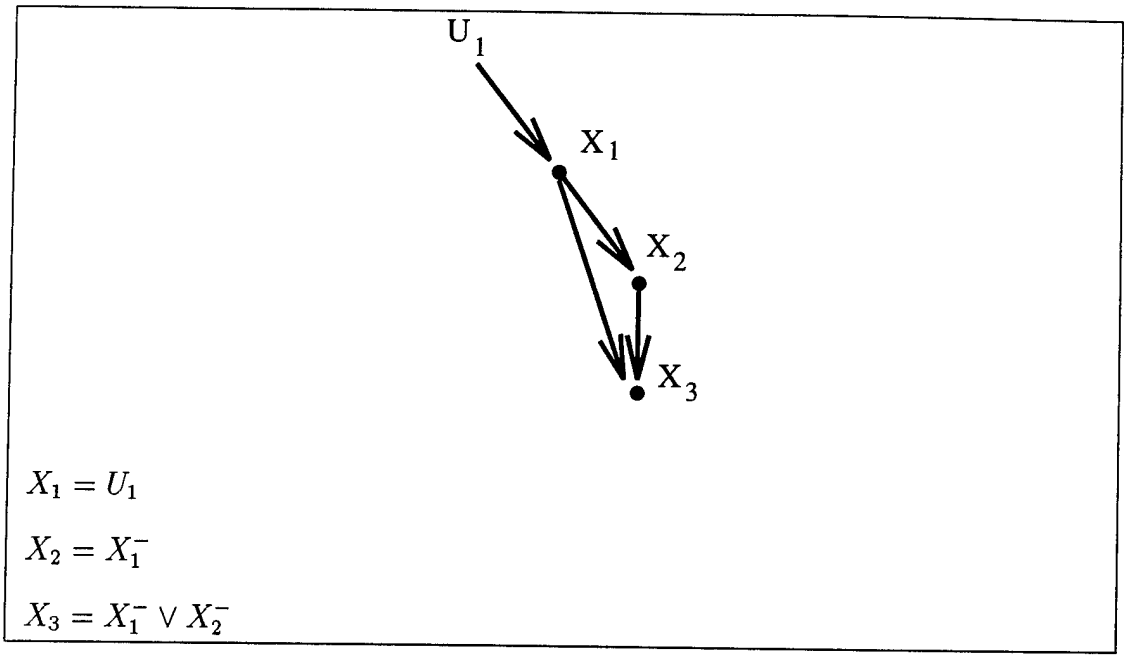


Figure 3.3: A causal model with memory which contains variables whose values depend upon the past history of other variables.

causal model. Given any arbitrary causal model with memory, we can create a new causal model with memory such that for all $X_i \in V$, $\mathbf{pa}^-_i = X_i^-$, while maintaining the information on how the model will behave under interventions. We do this by increasing the domain of each of the variables to include previous state information.

For example, consider the causal model in Figure 3.3, which contains three binary variables. The functions for X_2 and X_3 both rely on the past values of other variables. We can create an equivalent causal model with memory such that each variable depends only upon its own past value by making all the variables have four values instead of two, as in figure 3.4. In this model, the high-order bit of each variable encodes the "past" value of the variable in the origional model, and the low-order bit encodes the "current" value of the variable in the origional model. In the functions in example, *MOD* stands for modular division

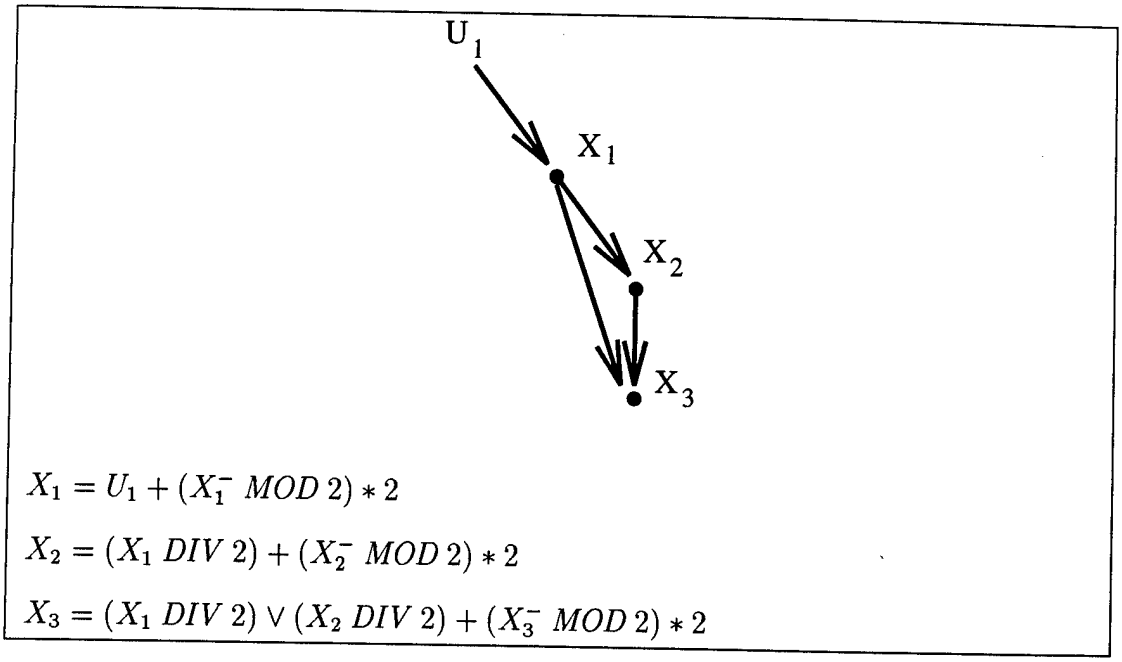


Figure 3.4: A causal model with memory which contains variables whose values do not depend upon the past histories of other variables

(so $X \ MOD \ 2$ extracts the low-order bit of X), and DIV stands for integral division (so $X \ DIV \ 2$ extracts the high-order bit of X). Likewise, multiplying by two is the same as a left shift.

We can make this transformation for any causal model with memory. The number of bits required to represent each variable may need to double, but the basic causal structure of the new model will be the same as that of the origional model. Thus, limiting $\mathbf{pa}^-_i$ to X_i^- does not limit the expressive power of a causal model with memory.

3.3 Time-Series Causal Model

A logical extension of adding a single piece of memory to a causal model is to consider a time series. We can time-index each of the variables in V and U. Thus,

V will stand for a time series, $V_{t=1}, V_{t=2}, \ldots$. Likewise, U will be a time series $U_{t=1}, U_{t=2}, \ldots$. Each function f_i will map a subset of $V_{t-1} \cup U_t \cup V_t$ to X_{it}. Thus, we will be able to consider how a system changes over time. Formally:

Definition 16 (time-series causal model) *A* time-series causal model *is a* 3*-tuple*

$$M = \langle V, U, F \rangle$$

where

(i) $V = \{V_{t=1}, V_{t=2}, \ldots\}$ *is a time series of variables, where each* $V_t = \{X_{1t}, X_{2t}, \ldots, X_{nt}\}$ *is a set of endogenous variables determined within the system,*

(ii) $U = \{U_{t=1}, U_{t=2}, \ldots\}$ *is a time series of variables, where each* $U_t = \{U_{1t}, U_{2t}, \ldots U_{mt}\}$ *is a set of exogenous variables that represent disturbances, abnormalities, assumptions, or boundary conditions, and*

(iii) $F = \{f_i\}$ *is a set of* n *deterministic, nontrivial functions, each of the form*

$$X_i = f_i(\mathbf{pa}_{it}, \mathbf{pa}_{i(t-1)}, u) \quad i = 1, \ldots, n \quad \mathbf{pa}_{it} \subseteq V_t \setminus X_i, \mathbf{pa}^{\mathbf{i}}_i \subseteq V_{t-1}$$

The members of the set PA_i *(connoting parents) are often called the direct causes of* X_i*. We will assume that each variable has the same parent set over time. In addition, we often will make the assumption that* $P(u_t) = P(u_{t+1})$ *for all* t*; this assumption is not necessary, however. We will also assume that the set of equations in (iii) has a unique solution for* $X_{1t}, \ldots, X_{nt}$ *given any values of the disturbances* $U_{1(t)}, \ldots, U_{m(t)}$ *and previous values* V_{t-1}*. Thus, we can consider each variable* Y *to be a function of the disturbances* U *and the previous values of* V *in the causal model* M*:* $Y_t = Y_{M_t}(u_t, v_{t-1})$*.*

We can define probabilistic time-series causal model as follows.

Definition 17 (probabilistic time-series causal model) *A* probabilistic time-series causal model *is a tuple*

$$\langle M, P(u) \rangle$$

where

(i) M *is a time-series causal model* $\langle V, U, F \rangle$*, and*

(ii) $P(u)$ *is a probability distribution over* U *such that each element* $U_i \in U$ *is marginally independent of all other elements of* U*.*

Time-series causal models bear a strong resemblance to dynamic systems, particularly the area of discrete iterations. However, there are some significant differences.

Consider the definition of discrete iterations as defined by Robert [Rob86a]:

- X denotes a (usually large) finite set of variables
- F denotes a map of X onto itself
- x^0 denotes an initial value for X

Starting with the initial value x^0 of X, the model deals with the sequence of values defined by

$$X^{r+1} = F(x^r) \qquad r = 0, 1, 2, \ldots$$

Since X is finite, the sequence must converge to a value ξ, such that either $F(\xi) = \xi$, or converge to a cycle $\xi_0, \ldots, \xi_p$ such that

$$\begin{aligned}
\xi_2 &= F(\xi_1) \\
&\vdots \\
\xi_p &= F(\xi_{p-1}) \\
\xi_1 &= F(\xi_p)
\end{aligned}$$

In the model, there can be a single fixed point ξ such that for any value x^0 of X there exists a p such that for all $i > p$, $x^i = \xi$ or a single cycle $\xi_1 \ldots \xi_p$. Likewise, the model could inculde several cycles, fixed points, or a mixture of cycles and fixed points.

At first glance, the restriction on time-series causal models that each set of equations F must have a unique value x_t for each value u of U and x_{t-1} of X_{t-1} is overly strong. Indeed, at first glance, this restriction seems similar to insisting that a discrete iteration system have a single fixed point. However, restriction is actually more akin to requiring that the function F in the theory of discrete iteration systems have a unique value for every element x of X.

Consider for a moment what relaxing the restriction would mean for causal models. Given a set of boundary conditions and the previous value for each variable, there would be multiple possible values for the next time step. This would correspond to a nondeterministic world in which the nondeterminism could not be fully characterized or described. Now, consider how a similar situation plays out with the restriction in place. Let x_{t-1} and u be the values for which there could be more than one value of x_t. We can add an additional variable U_j to U which determines which of the possible values x_t will result from u and x_{t-1}. With the restriction, then, We can describe the behavior of the system

more completely that we could with the restriction relaxed; we do not have to leave part of the causal mechanism undefined.

This example illustrates one of the significant differences between time-series causal models and discrete iteration systems, that is, a time-series causal model has boundary conditions, expressed by the exogenous variables U, while a discrete iteration system does not

3.4 Conclusion

In this chapter, we proposed causal models with memory, an extension to causal models that implements the concept of previous state. This extension allow us to expand the range of systems that can be modeled using causal models. We also proposed a time-series causal model. This model acts as a bridge between full dynamic systems an causal models: it combines the expressiveness of the former with the compactness and ease of computation of the latter. Thus, the benifits of each formalism can be combined into one system. Consider modeling an economic system, where the researcher is interested in how changing interest rates over time will effect the price of a commodity. If the changes in interest rate are slow compared to the fluctuations of the price and quantity of the commodity, then the researcher will be able to combine the standard static equilibrium supply and demand equations with more dynamic elements. Thus, the researcher needs only model the dynamic behavior of the elements for which the transient effects are important, without giving up the power to model such dynamic systems.

CHAPTER 4

Causal Relevance

4.1 Introduction

Geiger, Verma, and Pearl [GVP90a] have developed a set of axioms for a class of relations called *graphoids*. These axioms characterize informational relevance[1] among observed events based on the semantics of conditional independence in probability calculus. This chaper develops a parallel set of axioms for *causal relevance*, that is, the tendency of certain events to affect the occurrence of other events in the physical world, independent of the observer-reasoner. Informational irrelevance is concerned with statements of the form "X is conditionally independent of Y given Z," which means that, given the value of Z, gaining information about X gives us no new information about Y. Causal irrelevance is concerned with statements of the form "X is causally irrelevant to Y given Z," which we take to mean "Changing X will not alter the value of Y, if Z is fixed."

The notion of causal relevance has its roots in the philosophical works of Good [Goo61], Suppes [Sup70] and Salmon [Sal84]. They have attempted to give probabilistic interpretations to cause-effect relationships, and to distinguish causal from statistical relevance. Although their attempts have not produced an algo-

[1]The term "relevance" will be used primarily as a generic name for the relationship of being relevant or irrelevant. It will be clear from the context when "relevance" is intended to negate "irrelevance."

rithmic definition of causal relevance, they did generate methods for testing the consistency of relevance statements against a given probability distribution and a given temporal ordering among the variables [Car89, Eel91, Pea96b]. This chapter aims at axiomatizing relevance statements in themselves, without reference to underlying probabilities or temporal orderings.

Axiomatic characterization of causal relevance may serve as a normative standard for theories of action as well as a guide for developing representation schemes (e.g., graphical models) for planning and decision-making applications. For example, instead of explicitly storing all possible effects of an action, as in STRIPS [FN72], such representation schemes should enable an agent to examine only direct effects of actions and to infer which actions are relevant for a given goal, and which cease to be relevant once other actions are implemented.

An axiomization of causal relevance could also be useful to experimental researchers in domains where exact causal models do not exist. If we know, through experimentation, that some variables have no causal influence on others in a system, we may wish to determine whether other variables can gain such causal influence under different experimental conditions, or we may want to discover what additional experiments could provide such information. For example, suppose we find that a rat's diet has no effect on tumor growth while the amount of exercise is kept constant and, conversely, that exercise has no effect on tumor growth while diet is kept constant. We would like to be able to infer that controlling only diet (while paying no attention to exercise) would still have no influence on tumor growth. A more subtle inference problem is whether changing cage temperature could have an effect on the rat's physical activity, having established that temperature has no effect on activity when diet is kept constant

and that temperature has no effect on (the rat's choice of) diet when activity is kept constant.

We provide two formal definitions of causal irrelevance, a probabilistic definition and a deterministic definition. The probabilistic definition, which equates causal irrelevance with inability to change the probability of the effect variable, has intuitive appeal but is inferentially very weak; it does not support a very expressive set of axioms unless further assumptions are made about the underlying causal model. If we add the stability assumption (i.e., that no irrelevance can be destroyed by changing the nature of the individual processes in the system), then we obtain a set of axioms for probabilistic causal irrelevance that is the sane as the set governing path-interception in directed graphs. The deterministic definition, which equates causal irrelevance with inability to change the effect variable (in any state of the world), allows for a richer set of axioms without any assumptions about the causal model being made. All of the path-interception axioms for directed graphs, with the exception of transitivity, hold for deterministic causal irrelevance.

4.2 Probabilistic Causal Irrelevance

The existence of a probability distribution over all of the variables in a causal model (Eq. (2.4)) leads to a natural definition of a probabilistic version of causal irrelevance.

Definition 18 (probabilistic causal irrelevance) X *is* probabilistically causally irrelevant *to* Y, *given* Z, *written* $(X \not\rightarrow Y|Z)_P$, *iff*

$$\forall x, x', y, z \quad P(y|\hat{z}, \hat{x}) = P(y|\hat{z}, \hat{x}')$$

Read: "Once we hold Z fixed (at z), changing X between any two values will not affect the probability of Y."

4.2.1 Comparison to Informational Relevance

If we remove the "hats" from Definition 18, we get the standard definition of conditional independence in probability calculus, denoted $(X \perp Y|Z)$, which is governed by the graphoid axioms [PP87, GVP90b] given in Figure 4.1. Dawid [Daw79] and Spohn [Spo80] introduced different forms of these axioms, and Pearl and Paz [PP87] conjectured that these axioms were complete. Studeny [Stu92] refuted This conjecture and proved that conditional independence in probability theory has no finite axiomatization. Nevertheless, the graphoid axioms capture the most important features of informational relevance: "Learning irrelevant information should not alter the relevance status of other propositions in the system; what was relevant remains relevant, and what was irrelevant remains irrelevant" [Pea88].

One of the salient differences between informational and causal relevance is the property of symmetry, axiom 1.1. Informational relevance is symmetric, namely, if X is relevant to Y, then Y is relevant to X as well. For example, learning whether the sprinkler is on provides information on whether the pavement is wet, and, vice versa, learning whether the pavement is wet provides information on whether the sprinkler is on. This property is clearly violated in causal models: turning a sprinkler on tends to make the pavement wet, so turning on the sprinkler gives us information about the state of the pavement; conversely, wetting the

1.1 (Symmetry) $(X \perp Y|Z) \Longrightarrow (Y \perp X|Z)$

1.2 (Decomposition) $(X \perp YW|Z) \Longrightarrow (X \perp Y|Z)$

1.3 (Weak Union) $(X \perp YW|Z) \Longrightarrow (X \perp Y|ZW)$

1.4 (Contraction) $(X \perp Y|Z)\ \&\ (X \perp W|ZY) \Longrightarrow (X \perp YW|Z)$

1.5 (Intersection) $(X \perp W|ZY)\ \&\ (X \perp Y|ZW) \Longrightarrow (X \perp YW|Z)$

Intersection requires a strictly positive probability distribution.

Figure 4.1: The graphoid axioms

pavement has no physical effect on the state of the sprinkler and gives us no information about whether the sprinkler was on or off.

Another basic difference between informational and causal relevance is that in the former, the rule of the hypothetical middle [Pea88, p. 17] always holds:

$$\text{MIN}_x P(y|x) \leq P(y) \leq \text{MAX}_x P(y|x) \tag{4.1}$$

In causal relevance, $P(y)$ might be greater than $\text{MAX}_x P(y|\hat{x})$ or less than $\text{MIN}_x P(y|\hat{x})$. Figure 4.2 illustrates such a possibility.

In Figure 4.2, there are two endogenous variables X and Y, as well as an exogenous variable U_1. Without any intervention, X will always have the same value as U_1, and thus, Y will have the value 1. If X and U_1 have different values, however, then Y will have the value 0. If we intervene and set $X = 1$, then Y will have the value 1 when $U_1 = 1$, which has a probability 0.5, and Y will have the value 0 when $U_1 = 0$, which has a probability 0.5: $P(Y = 0|set(X = 1)) = P(Y = 1|set(X = 1)) = 0.5$. Similarly, we can see that $P(Y = 0|set(X = 0)) = P(Y = 1|set(X = 0)) = 0.5$. Thus, $MAX_x P(y|\hat{x}) = 0.5$,

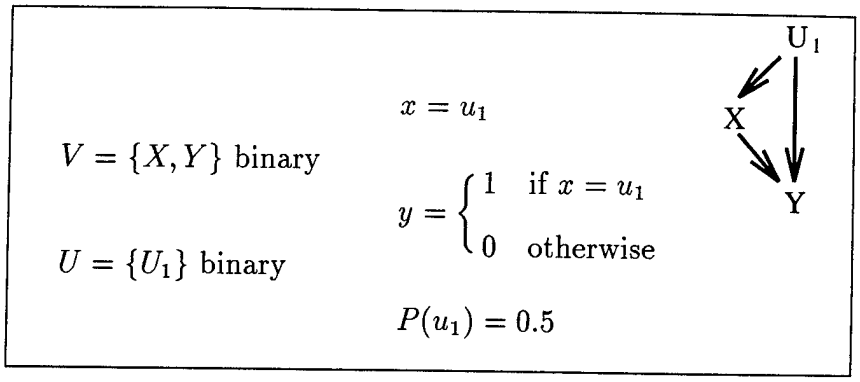


Figure 4.2: Example of $P(y) > \text{MAX}_x P(y|\hat{x})$

and $P(Y = 1) = 1 > 0.5 = \text{MAX}_x P(y|\hat{x})$.

Note that, given this violation of the rule of the hypothetical middle (Eq. (4.1)), Definition 18 is not equivalent to

$$\forall x, y, z \quad P(y|\hat{z}, \hat{x}) = P(y|\hat{z}) \tag{4.2}$$

Read: "Once we hold Z fixed (at z), controlling X will not affect the probability of Y." In fact, Eq. (4.2) is stronger than Definition 18; furthermore, statement 2.5.2 (left-intersection of Theorem 6, below) follows from the former but not from the latter.

The notion of probabilistic causal irrelevance may bring to mind the concept *ignorability* [RR83] which is extremely important in analyzing the effectiveness of treatments (e.g., drugs, diet, educational programs) from uncontrolled studies. The two concepts are related but different. Ignorability allows us to ignore *how* X obtained its value x, while irrelevance allows us to ignore *which* value X actually obtained. Ignorability is defined as the condition

$$P(Y_x = y|z) = P(Y = y|z, x) \tag{4.3}$$

which implies

$$P(y|\hat{x}) \doteq P(Y_x = y) = E_z(y|z,x) \tag{4.4}$$

Thus, ignorability allows an investigator to relate the potential response Y_x to observable conditional probabilities. Central in experimental design is the question of how to select a set of observables Z that would make Eq. (4.3) true, given causal knowledge of the domain. Ignorability in itself does not provide such a criterion, although it does state the problem in formal counterfactual language: "Z can be selected if, for every x, the value that Y would obtain had X been x is conditionally independent of X, given Z." A criterion for selecting Z can be obtained from the graph $G(M)$ underlying a causal model (e.g., the back-door criterion in [Pea95a]).

The question we attempt to answer in this section is whether the relation of causal irrelevance, $(A \not\rightarrow B|C)_P$, is governed by a set of axioms similar to the set of axioms governing the relation of informational irrelevance, $(A \perp B|C)$. More generally, one may ask whether there are any constraints that prohibit the assignment of arbitrary functions $P(y|\hat{x})$ to any pair (X,Y) of variable sets in V, in total disregard of the fact that $P(y|\hat{x})$ represents the probability of $(Y = y)$ induced by physically setting $X = x$ in some causal model M. Our finding indicate that, although it is not totally arbitrary, the assignment $P(y|\hat{x})$ is only weakly constrained by qualitative axioms such as those in Figure 4.1.

4.2.2 Axioms of Probabilistic Causal Irrelevance

We have found only two qualitative properties that constrain probabilistic causal irrelevance.

Theorem 6 *For any causal model, the following two properties must hold:*

2.2.1 (Right-Decomposition) $(X \not\to YW|Z)_P \Longrightarrow (X \not\to Y|Z)_P \,\&\, (X \not\to W|Z)_P$

2.5.2 (Left-Intersection) $(X \not\to Y|ZW)_P \;\&(W \not\to Y|ZX)_P \Longrightarrow (XW \not\to Y|Z)_P$

Property 2.2.1 reads: "If changing X has no effect on Y and W considered jointly, then it has no effect on either Y or W considered separately." This follows trivially from the fact that $P(\cdot)$ is a probability function, but it does not reflect any quality of causation.

Property 2.5.2 reads: "If changing X cannot affect $P(y)$ when W is fixed, and changing W cannot affect $P(y)$ when X is fixed, then changing X and W together cannot affect $P(y)$."

Many seemingly intuitive properties do not hold. For instance, none of the following statements holds for all causal models.

2.2.2 (Left-Decomposition-1) $(XW \not\to Y|Z)_P \Longrightarrow (X \not\to Y|Z)_P \vee (W \not\to Y|Z)_P$

2.2.3 (Left-Decomposition-2) $(XW \not\to Y|Z)_P \Longrightarrow (X \not\to Y|Z)_P \vee (X \not\to W|Z)_P$

2.2.4 (Left-Decomposition-3)

$(XW \not\to Y|Z)_P \;\wedge\; (XY \not\to W|Z)_P \Longrightarrow (X \not\to Y|Z)_P \vee (X \not\to W|Z)_P$

2.3 (Weak Union) $(X \not\to WY|Z)_P \Longrightarrow (X \not\to Y|ZW)_P$

2.4 (Contraction) $(X \not\to Y|Z)_P \;\wedge\; (X \not\to W|ZY)_P \Longrightarrow (X \not\to WY|Z)_P$

2.5.1 (Right-Intersection) $(X \not\to Y|ZW)_P \wedge (X \not\to W|ZY)_P \Longrightarrow (X \not\to WY|Z)_P$

2.6 (Transitivity) $(X \not\to Y|Z)_P \Longrightarrow (a \not\to Y|Z)_P \vee (X \not\to a|Z)_P \quad \forall a \notin X \cup Z \cup Y$

The sentences above were tailored after the graphoid axioms (Figure 4.1) with the provision that symmetry does not hold, thus requiring left and right versions.

Many of these sentences have intuitive appeal and yet are not sound relative to the semantics of $P(y|\hat{x})$. For example; left-decomposition states that if changing X has an effect on Y, and changing W has an effect on Y, then changing X and W simultaneously should also affect Y. It is hard to find a simple real-life example that refutes this assertion. Still, as will be shown in the examples of Section 4.3.1 and in Appendix A, each of these sentences is refuted by some specific causal model.

4.3 Proofs of Axioms of Probabilistic Causal Irrelevance

We now prove the two sentences of Theorem 6.

2.2.1 Holds trivially. $(X \not\rightarrow YW|Z)_P \Longrightarrow P(yw|\hat{z},\hat{x}) = P(yw|\hat{z})$. We can sum over W to get $P(y|\hat{z},\hat{x}) = P(y|\hat{z})$, which implies $(X \not\rightarrow Y|Z)_P$. The case for $(X \not\rightarrow W|Z)_P$ is symmetric □

2.5.2 (By Contradiction) Assume $(X \not\rightarrow Y|ZW)_P \wedge (W \not\rightarrow Y|ZX)_P \wedge \neg(XW \not\rightarrow Y|Z)_P$. Since $\neg(XW \not\rightarrow Y|Z)_P$, by definition $\exists y, x, x', w, w', z \;\; P(y|\hat{z},\hat{w},\hat{x}) \neq P(y|\hat{z},\hat{w}',\hat{x}')$. However, $(X \not\rightarrow Y|ZW)_P$ implies $\forall y, x, x', z, w \; P(y|\hat{z},\hat{x},\hat{w}) = P(y|\hat{z},\hat{x}',\hat{w})$. Furthermore, $(W \not\rightarrow Y|ZX)_P$ implies $\forall y, x', w, w', z \; P(y|\hat{z},\hat{x}',\hat{w}) = P(y|\hat{z},\hat{x}',\hat{w}')$ So, $\forall x, x', w, w', z \;\; P(y|\hat{z},\hat{x},\hat{w}) = P(y|\hat{z},\hat{x}',\hat{w}) = P(y|\hat{z},\hat{x}',\hat{w}')$. Thus $\forall x, x', w, w', z \; P(y|\hat{z},\hat{x},\hat{w}) = P(y|\hat{z},\hat{x}',\hat{w}')$, which contradicts $\exists x, x', w, w', z$ $P(y|\hat{z},\hat{x},\hat{w}) \neq P(y|\hat{z},\hat{x}',\hat{w}')$. □

4.3.1 Counterexample to Property 2.2.2

We now disprove property 2.2.2 by counterexample. This counterexample is not necessarily meant to model a common, real-life situation. Rather, it disproves the claim that *all possible* causal models conform to the property.

2.2.2 $(XW \not\rightarrow Y|Z)_P \Longrightarrow (X \not\rightarrow Y|Z)_P \vee (W \not\rightarrow Y|Z)_P$

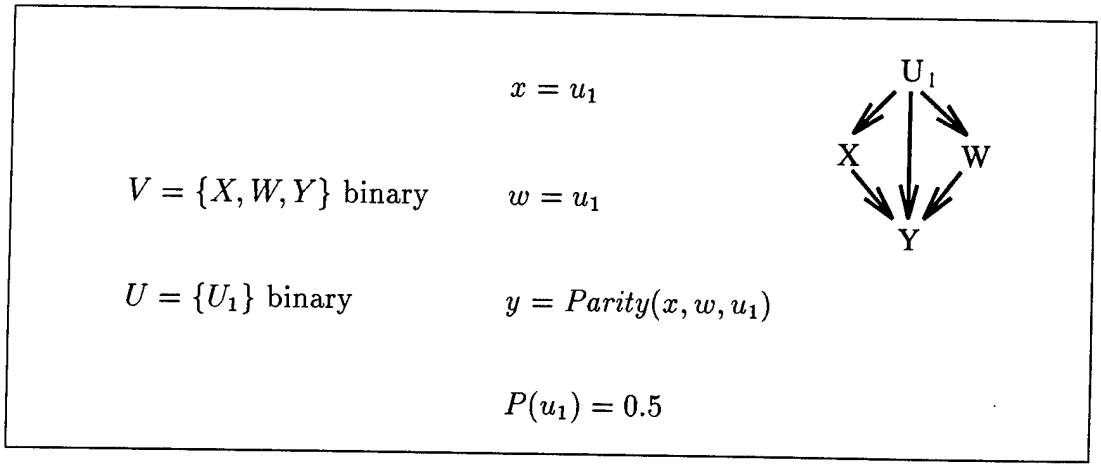


Figure 4.3: Counterexample to property 2.2.2.

Figure 4.3 shows a counterexample to this sentence. In this model, $(XW \not\rightarrow Y|\emptyset)_P \,\&\, \neg(X \not\rightarrow Y|\emptyset)_P \,\&\, \neg(W \not\rightarrow Y|\emptyset)_P$. The contrapositice form of this counterexample states that changing W can affect the probability of Y, and changing X can affect the probability of Y, but changing W and X simultaneously has no effect on the probability of Y. This is extremely counterintuitive; if tweaking X has an effect on Y, and tweaking W has an effect on Y, we would expect the more flexible option of changing X and W simultaneously to also affect Y. The key to this counterexample is the fact that setting W removes the connection between W and U_1. When we intervene on only X, W takes on the same value as U_1, and Y will always

have the value of X. When we intervene on both X and W, there is no longer any connection between U_1 and W. Thus, the probability that W and U_1 will have the same value is 0.5, and $P(y) = 0.5$

Counterexamples to the other six properties that do not hold for all causal models are in Appendix A.

4.3.2 Numeric Constraints

Although Definition 18 imposes only weak constraints (axioms 2.2.1 and 2.5.2) on the structure of probabilistic causal irrelevance, the probability assignments $P(y|\hat{x})$, which describe the effects of actions in the domain, are constrained nevertheless by nontrivial numerical bounds. For instance, the inequality

$$(y|\hat{x},\hat{z}) \geq P(y,z|\hat{x}) \tag{4.5}$$

must hold in any causal model. This can easily be shown by the definitions of $P(y,z|\hat{x})$ and $P(y|\hat{x},\hat{z})$. Recall from Eq. (2.4) that

$$P(y,z|\hat{x}) = \sum_{\{u \mid Y_x(u)=y \ \&\ Z_x(u)=z\}} P(u)$$

and

$$P(y|\hat{x},\hat{z}) = \sum_{\{u \mid Y_{xz}(u)=y\}} P(u)$$

Consider U^{yz}, the set of all values u of U such that $Y_x(u) = Y$ and $Z_x(u) = z$, and U^y_z, the set of all values u' of U such that $Y_{xz}(u') = y$. Since all values u of U^{yz} already constrain Z to have the value z, fixing Z at z will not affect the value of Y. Thus, for all values u of U^{yz}, $Y_{xz}(u) = y$. Hence, $U^y_z \supseteq U^{yz}$ and $P(y|\hat{x},\hat{z}) \geq P(yz|\hat{x})$. This can be shown more formally using Theorem 1 (Theorem 1 is proven in Section 2.7). Additional constraints were explored in [Pea95b].

4.3.3 Axioms of Causal Relevance for Stable Models

The set of axioms we obtained for causal irrelevance is much smaller than we would expect from our intuition about cause-effect relations. We have two explanations for this discrepancy. One possibility is that our intuition of causal relevance is based on a deterministic rather than a probabilistic conception of physical reality. This possibility will be explored in Section 4.4, where we give a deterministic definition of causal irrelevance that yields a more complete set of axioms. The other possibility is that the type of examples exploited in Section 4.3.1 and Appendix A are not commonly observed in everyday life. In this section, we explore what assumptions need to be made for probabilistic causal irrelevance to acquire properties that we intuitively associate with causal irrelevance.

A more expressive set of causal relevance axioms is obtained if we confine the analysis to *stable* causal models, that is, to causal models whose irrelevances are implied by the structure of the causal model and, hence, remain invariant to changes in the forms of each individual function f_i. Our definition of stability employs the concept of a replacement class. A *replacement class* τ is the set of all models that have the same variables V and U, and the same functional arguments. In other words, the functions are allowed to change between members of τ, but the arguments of these functions are not allowed to vary. Formally, for any two models $M_1, M_2 \in \tau$ and any two functions $f_i(PA_i) \in M_1$ and $f'_i(PA'_i) \in M_2$, $PA_i = PA'_i$. The class $\tau(M)$ represents the replacement class that contains the model M.

We now define stability using replacement classes (see also [PV91][2]).

[2]The probabilistic notion of stability (also called "DAG-isomorphism," "nondegeneracy" [Pea88, p. 391], and "faithfulness" [SGS93]) was used by Pearl and Verma [1991] to emphasize the invariance of certain independencies to functional form.

Definition 19 (stability) *Let M be a causal model. An irrelevance $(X \not\rightarrow Y|Z)_P$ in M is* stable *if it is shared by all models in $\tau(M)$. The model M is* stable *if all of the irrelevances in M are stable.*

Stability requires that irrelevance be determined by the structure of the equations, not merely by the parameters of the functions. Thus, a causal model is not stable if we can remove an irrelevance relationship by replacing an equation or set of equations to obtain a new model with fewer irrelevance statements. In each of the examples in Section 4.3.1 and Appendix A, for instance, a minor change in the form of one of the equations would destroy an irrelevance. None of the models presented in Figure 4.3 and the appendix is stable.

There are, however, many stable causal models. All monotonic linear systems, for example, are stable. One might think that any causal model that contains only additive, monotonic functions f_i would be stable. The causal model of Figure A.7 refutes that conjecture.

Definition 20 (path-interception) *Let $(X \twoheadrightarrow Y|Z)_G$ stand for the statement "Every directed path from X to Y in graph G contains at least one element in Z."*

Theorem 7 *If a causal model M is stable, then X is probabilistically causally irrelevant to Y, given Z, in M iff Z intercepts all directed paths from X to Y in the graph $G(M)$ defined by M. That is,*

$$(X \not\rightarrow Y|Z)_P \Longleftrightarrow (X \twoheadrightarrow Y|Z)_{G(M)}$$

Proof:

(i) $\quad (X \not\rightarrow Y|Z)_P \Longrightarrow (X \twoheadrightarrow Y|Z)_{G(M)}$

Assume that there exists a stable causal model M that induces a probabilistic causal irrelevance relation $(A \not\rightarrow B|C)_P$, and assume that, for some sets of variables X, Y, Z, $(X \not\rightarrow Y|Z)_P$ and $\neg(X \nrightarrow Y|Z)_{G(M)}$. Since there is a directed path from X to Y that is not intercepted by Z in $G(M)$, we can easily construct a model M' such that $G(M') = G(M)$ and $\neg(X \not\rightarrow Y|Z)_P$ in M'. We can do this by changing all of the functions that lie on the path from X to Y to disjunctions and then modifying the other functions to ensure that $P(y|\hat{z}) < 1$. Thus, if we force X to have the value 1, Y will also have the value 1, and $P(y|\hat{z}, \hat{x}) \neq P(y|\hat{z})$. By assumption, $(X \not\rightarrow Y|Z)_P$, so an irrelevance in M is not shared in a member of $\tau(M)$. Thus, M is not a stable causal model, a contradiction.

(ii) $\qquad (X \nrightarrow Y|Z)_{G(M)} \Longrightarrow (X \not\rightarrow Y|Z)_P$

We will use the following lemma:

Lemma 1 *For any structural equation f_Y in a causal model M, if a series of functional substitutions results in a new function g_Y such that X is an argument of g_Y, then there must be a directed path from X to Y in $G(M)$.*

We will prove this lemma by induction on the number of functional substitutions.

Base Case: If we make no substitutions into f_Y, then every argument X of f_Y must be a parent of Y in $G(M)$, by our definition of $G(M)$. Thus, there is a directed path from each argument of f_Y to Y in $G(M)$

Inductive Case: Assume that $n-1$ functional substitutions into f_Y always results in the new function g_Y such that for each argument X of g_Y, there is a directed path from X to Y in $G(M)$. We use this assumption to prove that after n substitutions resulting in g'_Y, there is a directed path from every argument of g'_Y to Y in $G(M)$, as follows: When we do a single substitution, we replace a

variable with a function of its parents in $G(M)$. So, for any new argument X' that is introduced into g'_Y by substituting in for X, X' must be a parent of X in $G(M)$. By the inductive hypothesis, there must be a directed path from X to Y in $G(M)$. Thus, there must be a directed path from X' to Y in $G(M)$.

We can now prove the implication $(X \twoheadrightarrow Y|Z)_{G(M)} \Longrightarrow (X \not\twoheadrightarrow Y|Z)_P$. We will consider f_Y, the functional equation for Y in M_z. After we do a functional substitution for all variables in f_Y except X and Z, we are left with a new function g_Y. By Lemma 1, since there is no directed path from X to Y in $G(M_z)$, X is not an argument of g_Y, so g_Y is a function of only Z and U. Since g_Y is a function of only Z and U, and not of X, $Y_{xz}(u) = Y_z(u)$; hence, $P(y|\hat{x},\hat{z}) = P(y|\hat{z})$, and $(X \not\twoheadrightarrow Y|Z)_P$. □

Since $(X \not\twoheadrightarrow Y|Z)_P \Longleftrightarrow (X \twoheadrightarrow Y|Z)_{G(M)}$ in stable causal models, probabilistic causal irrelevance is completely characterized by the axioms of path interception in directed graphs. A complete set of such axioms was developed in [PP94, PPU96] and is given in Figure 4.4.

3.2.1 (Right-Decomposition) $(X \twoheadrightarrow YW|Z)_G \Longrightarrow (X \twoheadrightarrow Y|Z)_G \,\&\, (X \twoheadrightarrow W|Z)_G$
3.2.2 (Left-Decomposition) $(XW \twoheadrightarrow Y|Z)_G \Longrightarrow (X \twoheadrightarrow Y|Z)_G \,\&\, (W \twoheadrightarrow Y|Z)_G$
3.4 (Strong Union) $(X \twoheadrightarrow Y|Z)_G \Longrightarrow (X \twoheadrightarrow Y|ZW)_G \quad \forall\, W$
3.5.1 (Right-Intersection) $(X \twoheadrightarrow Y|ZW)_G \,\&\, (X \twoheadrightarrow W|ZY)_G \Longrightarrow (X \twoheadrightarrow YW|Z)_G$
3.5.2 (Left-Intersection) $(X \twoheadrightarrow Y|ZW)_G \,\&\, (W \twoheadrightarrow Y|ZX)_G \Longrightarrow (XW \twoheadrightarrow Y|Z)_G$
3.6 (Transitivity) $(X \twoheadrightarrow Y|Z)_G \Longrightarrow (a \twoheadrightarrow Y|Z)_G \vee (X \twoheadrightarrow a|Z)_G \quad \forall a \notin X \cup Z \cup Y$

Figure 4.4: Sound and complete axioms for path-interception in directed graphs

4.4 Deterministic Causal Relevance

The notion of causal irrelevance obtains a deterministic definition when we consider the effects of an action conditioned on a specific state of the world u.

Definition 21 (causal irrelevance) X *is* causally irrelevant *to* Y, *given* Z, *written* $(X \not\rightarrow Y|Z)_T$, *if*

$$\forall u, z, x, x' \quad Y_{xz}(u) = Y_{x'z}(u) \tag{4.6}$$

in every submodel of M_z.

This definition captures the intuition "If X is causally irrelevant to Y, then X cannot affect Y under any circumstance." Note that, unlike the probabilistic definition of causal irrelevance (see Eq. (4.2)), the deterministic definition is equivalent to

$$\forall u, z, x \quad Y_{xz}(u) = Y_z(u) \tag{4.7}$$

Moreover,it is stronger than the probabilistic definition, in that $(X \not\rightarrow Y|Z)_T \Longrightarrow (X \not\rightarrow Y|Z)_P$.

This definition of irrelevance bears some similarity to the idea of limited unresponsiveness presented by Heckerman and Shacter [HS95]. However, whereas they define causality in terms of limited unresponsiveness to a specific set of actions, we view irrelevance as a property of the configuration of the mechanisms that compose causal model. In fact, a version of their definition of causality, translated into our language, will be shown to be a theorem of causal irrelevance in Section 4.4.4.2 (see Eq. (4.8)).

To see why we require the equality $Y_{xz}(u) = Y_{x'z}(u)$ to hold in every submodel of M_z, consider the causal model of Figure 4.5. In this example, Z follows X

and, hence, Y follows X, that is, $Y_{X=0}(u) = Y_{X=1}(u) = u_2$. However, since f_y is a nontrivial function of X, X is perceived to be causally relevant to Y. Only holding Z constant reveals the causal influence of X on Y. To capture this intuition, we must consider all submodels M_z in Definition 21.

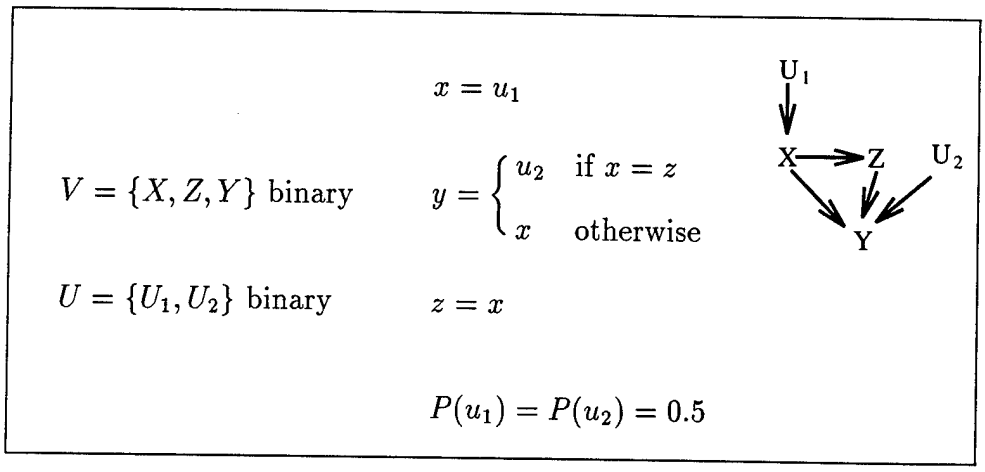


Figure 4.5: Example of a causal model that requires the examination of submodels before causal relevance can be determined

4.4.1 Axioms of Causal Irrelevance

Given this definition of causal irrelevance, we have the following theorem:

Theorem 8 *For any causal model, the following sentences must hold:*

4.2.1 (Right-Decomposition) $(X \not\rightarrow YW|Z)_T \Longrightarrow (X \not\rightarrow Y|Z)_T \,\&\, (X \not\rightarrow W|Z)_T$

4.2.2 (Left-Decomposition) $(XW \not\rightarrow Y|Z)_T \Longrightarrow (X \not\rightarrow Y|Z)_T \,\&\, (W \not\rightarrow Y|Z)_T$

4.4 (Strong Union) $(X \not\rightarrow Y|Z)_T \Longrightarrow (X \not\rightarrow Y|ZW)_T \;\; \forall\, W$

4.5.1 (Right-Intersection) $(X \not\rightarrow Y|ZW)_T \,\&\, (X \not\rightarrow W|ZY)_T \Longrightarrow (X \not\rightarrow YW|Z)_T$

4.5.2 (Left-Intersection) $(X \not\rightarrow Y|ZW)_T \,\&\, (W \not\rightarrow Y|ZX)_T \Longrightarrow (XW \not\rightarrow Y|Z)_T$

The following sentence, however, *does not* hold in every causal model:

4.6 (Transitivity) $(X \not\rightarrow Y|Z)_T \Longrightarrow (a \not\rightarrow Y|Z)_T \vee (X \not\rightarrow a|Z)_T \quad \forall a \notin X \cup Z \cup Y$

4.4.2 Proofs of Causal Irrelevance Axioms

Using the theorems from Section 2.7, we can prove Theorem 8, the axioms of causal relevance are sound.

4.2.1 Holds trivially. □

4.2.2 (By contradiction) Assume that there exists a causal model such that $(XW \not\rightarrow Y|Z)_T \;\&\; \neg((Z \not\rightarrow Y|Z)_T \;\&\; (W \not\rightarrow Y|Z)_T)$. So, either $(XW \not\rightarrow Y|Z)_T \& \neg(X \not\rightarrow Y|Z)_T$ or $(XW \not\rightarrow Y|Z)_T \& \neg(W \not\rightarrow Y|Z)_T$. First, we consider $(XW \not\rightarrow Y|Z)_T \,\&\, \neg(X \not\rightarrow Y|Z)_T$. By our definition of causal irrelevance, $\neg(X \not\rightarrow Y|Z)_T$ implies that there exist two values x, x' of X and some value u of U such that $Y_{xz}(u) \neq Y_{x'z}(u)$. Now, let us consider the values x, x', z, u such that $Y_{xz}(u) \neq Y_{x'z}(u)$. Using these values, we can determine w and w' as follows: Let $w = W_{xz}(u)$, and $w' = W_{x'z}(u)$. It does not matter whether $w = w'$ or $w \neq w'$. By composition, $Y_{xzw}(u) \neq Y_{x'zw}(u)$. Thus, $\exists x, w, z, u \;\; Y_{xwz}(u) \neq Y_{x'w'z}(u)$, which contradicts $(XW \not\rightarrow Y|Z)_T$. Thus, $(XW \not\rightarrow Y|Z)_T \,\&\, \neg(X \not\rightarrow Y|Z)_T$ leads to a contradiction. We can use a symmetric argument to show that $(XW \not\rightarrow Y|Z)_T \,\&\, \neg(W \not\rightarrow Y|Z)_T$ also leads to a contradiction. □

4.4 By our definition of causal irrelevance, $(X \not\rightarrow Y|Z)_T \Longrightarrow Y_{xz}(u) = Y_{x'z}(u)$ for all submodels of M_{xz}. For an arbitrary W, we consider the submodel

M_w where W is forced to have the value w. By our definition of causal irrelevance, $Y_{xzw}(u) = Y_{x'zw}$ for all values w. In addition, since $(X \not\rightarrow Y|Z)_T \Longrightarrow Y_{xz}(u) = Y_{x'z}(u)$ for all submodels of M, $Y_{xzw}(u) = Y_{x'zw}$ for all submodels of M_w. Since W was arbitrary, $(X \not\rightarrow Y|Z)_T \Longrightarrow (X \not\rightarrow Y|ZW)_T$ for all W. □

4.5.1 (By contradiction) Assume $(X \not\rightarrow Y|ZW)_T \& (X \not\rightarrow W|ZY)_T \& \neg(X \not\rightarrow YW|Z)_T$. $\neg(X \not\rightarrow YW|Z)_T$ implies $\exists x, x', z \ (Y_{xz}(u) \neq Y_{x'z}(u)) \vee (W_{xz}(u) \neq W_{x'z}(u))$. Since W and Y are symmetric, we will only consider Y. Consider the values of x, x', z, u such that $Y_{xz}(u) \neq Y_{x'z}(u)$. Let $y = Y_{xz}(u)$ and $y' = Y_{x'z}(u)$. By composition, $Y_{xz}(u) = Y_{xzw}(u)$ for $w = W_{xz}(u)$. By assumption, $Y_{xzw}(u) = Y_{x'zw}(u)$. Also by composition, $W_{xz}(u) = W_{xzy}(u)$ for $y = Y_{xz}(u)$. By assumption, $W_{xzy}(u) = W_{x'zy}(u)$. By reversibility, since y is a solution to the simultaneous equations $y = Y_{x'zw}$ and $w = W_{x'zy}$, then y must also be a solution to $Y_{x'z}(u)$. Thus $y = y'$, a contradiction. We can use a symmetric argument to show that $W_{xz}(u) \neq W_{x'z}(u)$ also leads to a contradiction. □

4.5.2 (By contradiction) Assume $(X \not\rightarrow Y|ZW)_T$ & $(W \not\rightarrow Y|ZX)_T$ & $\neg(XW \not\rightarrow Y|Z)_T$. Since $\neg(XW \not\rightarrow Y|Z)_T$, by definition $\exists x, x', w, w', z \ \ Y_{xwz}(u) \neq Y_{x'w'z}(u)$. However, $(X \not\rightarrow Y|ZW)_T$ implies $\forall x, x', z, w \ \ Y_{xzw}(u) = Y_{x'zw}(u)$. Furthermore, $(W \not\rightarrow Y|ZX)_T$ implies $\forall x', w, w', z \ \ Y_{x'wz}(u) = Y_{x'w'z}(u)$. Thus, $\forall x, x', w, w', z \ \ Y_{xwz}(u) = Y_{x'wz}(u) = Y_{x'w'z}(u)$, and so $\forall x, x', w, w', z \ \ Y_{xwz}(u) = Y_{x'w'z}(u)$. This contradicts $\exists x, x', w, w', z \ \ Y_{xwz}(u) \neq Y_{x'w'z}(u)$. □

4.4.3 Why Transitivity Fails in Causal Relevance

Causal transitivity is a property that makes intuitive sense. If a variable A has a causal influence on B, and B has a causal influence on C, one would think that A would have a causal influence on C. This is not always the case, however, even in deterministic causality. Consider the causal model described in Figure 4.6. In this example, X is causally relevant to W, and W is causally relevant to Y, but X is causally irrelevant to Y. The intuition behind this example is that changing X causes only a minor change in W, while Y only responds to large changes in W.

$V = \{X, W, Y\}$, $x, y \in \{0, 1\}$, $w \in \{0, 1, 2, 3\}$

$U = \{U_1, U_2\}$ $u_1, u_2 \in \{0, 1\}$

$w = x + 2 * u_2$

$x = u_1$

$y = (w > 1)$

$P(u_1) = P(u_2) = 0.5$

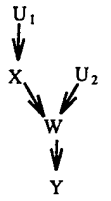


Figure 4.6: Counterexample to transitivity in causal irrelevance.

The failure of transitivity is deeper than this, however. Even when X has more complete control over the intermediate variable W, we still may not be able to achieve transitivity. Consider the causal model of Figure 4.7. This model is the same as the model of Figure 4.6 except W has now been split into $W_1, \ldots, W_4$, corresponding to W's four possible values. That is, W_1 is true if $x + u_2 = 0$, W_2 is true if $x + u_2 = 1$, W_3 is true if $x + u_2 = 2$, and W_4 is true if $x + u_2 = 3$. Now, by

fixing X, we can cause any of the intermediate variables $W_1, \ldots, W_4$ to be false in any given state of the world u. Likewise, each of the intermediate variables $W_1, \ldots, W_4$ can affect Y in any state u. However, X still has no effect on Y in any state u.

4.4.4 Causal Relevance and Directed Graphs

4.4.4.1 Causal Graphs as Irrelevance Maps

Comparing axioms 3.2–3.5 to axioms 4.2–4.5, we see that causal irrelevance is quite similar to path-interception in directed graphs. Since people (and machines) can easily reason about graphs, a graph that represents all of the causal relevances and irrelevances of a given causal model would be useful. That is, we would like to create a graph $G^*(M)$ such that

(i) Each variable X in M corresponds to exactly one node X^* in $G^*(M)$,

(ii) For all subsets of nodes X^*, Y^*, Z^* in $G^*(M)$, $(X^* \twoheadrightarrow Y^* | Z^*)_{G^*(M)} \Longrightarrow (X \not\rightarrow Y|Z)_T$, and

(iii) For all subsets of variables X, Y, Z in M, $(X \not\rightarrow Y|Z)_T \Longrightarrow (X^* \twoheadrightarrow Y^*|Z^*)_{G^*(M)}$.

In graph $G^*(M)$, if all directed paths from X^* to Y^* are intercepted by some variables in Z, then X is causally irrelevant to Y in the model M. Likewise, if a set of variables X is causally irrelevant to a set Y given fixed Z, then all paths from nodes in X^* to nodes in Y^* are intercepted by some variables in Z.

The obvious choice for $G^*(M)$ is $G(M)$, the graph associated with the causal model itself, as defined by Eq. (2.1). If we use $G^*(M) = G(M)$, then implication

(ii) holds, since in Section 4.3.3 we showed that $(X \twoheadrightarrow Y|Z)_{G(M)} \Longrightarrow Y_{xz}(u) = Y_z(u)$, and thus $(X \not\rightarrow Y|Z)_T$. However, since transitivity always holds in path interception but does not always in causal irrelevance, for a given model M there might be no graph $G^*(M)$ such that implications (ii) and (iii) hold simultaneously. Nonetheless, we can use directed graphs to validate candidate theorems of causal irrelevance, as we show next.

4.4.4.2 Directed Graphs as Theorem Provers

Consider an oracle that takes in statements about path-interception and returns YES if the statement holds in all directed graphs and NO otherwise. We will show that such an oracle can be used to validate or refute sentences about causal relevance.

First, let us consider a language of causal relevance in which the literals stand for simple irrelevance statements of the form $(X \not\rightarrow Y|Z)_T$, where X, Y and Z are sets of variables. Second, let the *canonical form* for sentences in the language of causal irrelevance be an implication $a_1 \,\&\, a_2 \,\&\, \ldots \,\&\, a_i \Longrightarrow b_1 \vee b_2 \vee \ldots \vee b_k$ whose antecedent consists of a conjunction of non-negated literals and whose consequent consists of non-negated literals. For instance, consider the sentence[3]

$$(X \not\rightarrow Y|Z)_T \,\&\, \neg(X \not\rightarrow Y|\emptyset)_T \Longrightarrow \neg(Z \not\rightarrow Y|\emptyset)_T \tag{4.8}$$

This sentence is not in canonical form because the second conjunct in the antecedent is negated and the statement in the consequent is negated. The canonical form of this sentence is

$$(X \not\rightarrow Y|Z)_T \,\&\, (Z \not\rightarrow Y|\emptyset)_T \Longrightarrow (X \not\rightarrow Y|\emptyset)_T \tag{4.9}$$

[3]A version of this sentence was chosen in [HS95] as the definition of causality.

Any causal irrelevance sentence can be written in a unique canonical form using standard logical procedures.

Definition 22 (Horn component) *A* Horn component *H of a causal irrelevance sentence S is a sentence H such that*

(i) H is in canonical form,

(ii) The consequent of H contains no disjunctions, and

(iii) $H \Longrightarrow S$.

If a sentence S is in the canonical form $a_1 \,\&\, a_2 \,\&\, \ldots \,\&\, a_i \Longrightarrow b_1 \vee b_2 \vee \ldots \vee b_k$, then a Horn component of S is any sentence of the form $a_1 \,\&\, a_2 \,\&\, \ldots \,\&\, a_i \Longrightarrow b_j$. For example, Eq. (4.9) has no disjunctions in its consequent and, hence, is itself a Horn component.

For any causal irrelevance statement A of the form $(X \not\rightarrow Y|Z)_T$, we will consider A_g, the *graphical translation* of A, to be the corresponding path-interception statement $(X \not\twoheadrightarrow Y|Z)_{G(M)}$. Using this convention, we can define

Theorem 9 (graphical theorem verification) *A causal irrelevance sentence S is true for all causal models iff there exists a Horn component H of S such that H_g, the graphical translation of H, is true for all graphs.*

For example, consider the sentence in Eq. (4.8). The canonical form of this sentence is given in Eq. (4.9) and is itself a Horn component. The sentence corresponding to Eq. (4.9) for path-interception in directed graphs, $(X \not\twoheadrightarrow Y|Z)_G \,\&\, (Z \not\twoheadrightarrow Y|\emptyset)_G \Longrightarrow (X \not\twoheadrightarrow Y|\emptyset)_G$, states that if all paths from X to Y are intercepted by Z, and there are no paths from Z to Y, then there is no path from X to Y.

This sentence is true for all directed graphs, so Eq. (4.8) is a valid theorem of causal relevance.

Next, consider transitivity, stated as $(X \not\to Y|Z)_T \Longrightarrow (a \not\to Y|Z)_T \vee (X \not\to a|Z)_T$. The Horn components of this sentence are

$$H^1 \quad : (X \not\to Y|Z)_T \Longrightarrow (a \not\to Y|Z)_T \tag{4.10}$$

$$H^2 \quad : (X \not\to Y|Z)_T \Longrightarrow (X \not\to a|Z)_T. \tag{4.11}$$

Looking at each of the corresponding path-interception sentences in turn, we find that neither $H^1_g : (X \not\twoheadrightarrow Y|Z)_G \Longrightarrow (a \not\twoheadrightarrow Y|Z)_G$ nor $H^2_g : (X \not\twoheadrightarrow Y|Z)_G \Longrightarrow (X \not\twoheadrightarrow a|Z)_G$ is true for all directed graphs G, that is, if Z intercepts all paths from X to Y, it is not the case that either Z intercepts all paths from any other variable to Y or Z intercepts all paths from X to any other variable. Thus, transitivity is not a theorem of causal relevance.

Proof (of Theorem 9):

First, we prove that if there are no disjunctions in the consequent of a canonical form sentence, then the sentence is true iff the corresponding sentence is true for path-interception in directed graphs.

We will prove this by contradiction. Assume that there exists some theorem $A \Longrightarrow B$, where A and B are conjunctions of literals such that

(i) $A \Longrightarrow B$ is not a theorem in causal irrelevance, and

(ii) $A_g \Longrightarrow B_g$ is a theorem in path-interception in directed graphs

Since $A_g \Longrightarrow B_g$ is a theorem in path-interception, then we must be able to generate B_g from A_g using the axioms of path-interception in directed graphs.

However, since $A \Longrightarrow B$ is not a theorem in causal irrelevance, every such generation of B_g from A_g must include application of the axiom of transitivity. When the axiom of transitivity is used, a disjunction is created. This disjunction must be used in the generation of B_g. By assumption, B_g does not contain a disjunction. Also, none of the antecedents of any of the axioms of path-interception contain disjunctions. Thus, the only way to use this disjunction in the generation of B_g is to resolve the disjunction with a negated clause. Since A_g started with no negated statements, and none of the axioms of path-interception can be used to create negated statements, we cannot resolve the disjunction with anything. Thus, generating B_g from A_g did not require an application of transitivity, a contradiction.

Next, we prove that if a theorem $A \Longrightarrow B \vee C$ is a theorem in causal irrelevance, then either $A \Longrightarrow B$ is a theorem in causal irrelevance or $A \Longrightarrow C$ is a theorem in causal irrelevance. If $A \Longrightarrow B \vee C$ is a theorem in causal irrelevance, then we must be able to generate $B \vee C$ from A using the axioms of causal irrelevance. Since no axiom creates a disjunction, to generate $B \vee C$ from A we must either generate B from A and add C or generate C from A and add B. Thus, a causal irrelevance sentence is a theorem iff there is a path-interception theorem that corresponds to one of the Horn components of the original sentence. □

4.5 Applications of Deterministic Causal Relevance

Frequently, researchers would like to know the causal effect of some variable, that is, $P(y|\hat{x})$. This question comes up especially in the medical sciences, where researchers want to determine the effectiveness of a particular drug or treatment. This quantity, $P(y|\hat{x})$, is often difficult, if not impossible, to measure. However,

the conditional probability $P(y|x)$ is often relatively easy to measure. We would like to be able to relate the measurable quantity, $P(y|x)$, to the desired quantity, $P(y\hat{x})$.

We can use the determinisitic definition of causal irrelevance to show when observation yields the same probability as action, that is, when $P(y|x) = P(y|\hat{x})$. The following theorem gives the conditions under which observation of X yields the same probability distribution on Y as intervention on X.

Theorem 10 *For any two variables X,Y, and for any two values x,y,* $P(y|\hat{x}) = P(y|x)$ *if* $\forall A \in U \cup V$, $(A \not\rightarrow X|\emptyset)_T \vee (A \not\rightarrow Y|X)_T$.

Theorem 10 is a precise statement of what is called *ignorability* in the statistical literature, and it justifies the use of randomized experiments to measure the quantity $P(y|\hat{x})$. For example, when trying to determine the causal effect of the dosage of a drug (X) on the the recovery of a patient (Y), researchers will often utilize a randomized experiment. The treatment assignment (Z) is randomized. Assuming perfect compliance, the dosage X each patient takes is determined completely by the treatment assignment Z, and the recovery Y is measured. Since Z is randomized, $\forall A, (A \not\rightarrow Z|\emptyset)_T$. If there is full compliance, then for any variable W other than Z, $(W \not\rightarrow X|\emptyset)_T$. If the experiment is double-blind, placebos are used to prevent the doctors and the patients from knowing the value of either the treatment assignment X or the dosage Z, and, thus, the treatment Z cannot affect the recovery Y except through the action of the drug. That is, $(Z \not\rightarrow Y|X)_T$. So, for any variable $A \in U \cup V$, if $A = Z$ then $(A \not\rightarrow Y|X)_T$, and $(W \not\rightarrow X|\emptyset)_T$ otherwise. Thus $\forall A \in U \cup V$, $(A \not\rightarrow X|\emptyset)_T \vee (A \not\rightarrow Y|X)_T$, and $P(y|\hat{x}) = P(y|x)$.

Proof of Theorem 10:

Given $\forall A \in U \cup V$, $(A \not\rightarrow X|\emptyset)_T \vee (A \not\rightarrow Y|X)_T$, we show that $P(y|\hat{x}) = P(y|x)$.

We separate the exogenous variables U into two groups, U' and U'', such that $(U' \not\rightarrow X|\emptyset)_T$ and $(U'' \not\rightarrow Y|X)_T$. We first consider the set of possible values u'' of U''. We are going to create a set of values u'' of U'', which we call α, that is the set of all values u'' of U'' for which there exists a value u' of U' such that $X(u', u'') = x$. Now we consider each element a of α, and for each element a we create a set β_a. Set β_a is the set of all values u' of U' such that $Y(u', a) = y$. We can now express $P(y|x)$ in terms of α and β_a:

$$P(y|x) = \frac{\sum_{a \in \alpha} P(a) \sum_{b \in \beta_a} P(b)}{\sum_{a \in \alpha} P(a)}$$

By the causal irrelevances that we used to separate U into U' and U'', and by composition, we know that for any two elements $a_1, a_2 \in \alpha$, $\beta_{a_1} = \beta_{a_2}$. We can thus define a new set γ such that $\forall a \in \alpha, \gamma = \beta_a$. Hence, the above formula can be simplified to

$$P(y|x) = \sum_{g \in \gamma} P(g)$$

We now consider the value of $P(y|\hat{x})$. By the irrelevancies that separate U into U' and U'', we know that, in the submodel with X fixed at x, Y is a function only of members of U', and not of members of U'', and

$$P(y|\hat{x}) = \sum_{u' \in U' | Y_x(u') = y} P(u')$$

We now consider the set $\{u' \in U' | Y_x(u') = y\}$. This set is the set of all values u' of U' such that when X has the value x, Y will have the value y. This set is identical to γ, so

$$P(y|\hat{x}) = \sum_{g \in \gamma} P(g) = P(y|x)$$

and $P(y|\hat{x}) = P(y|x)$. □

In other words, causal relevance gives a theoretical justification for the use of randomized experiments for measuring causal effects.

4.6 Conclusion

How do scientists predict the outcome of one experiment from the results of other experiments run under totally different conditions? Such transfer of experimental knowledge, although essential to scientific progress, involves inferences that cannot easily be formalized in the standard languages of logic, physics, or probability.

The formalization of such inferences requires a language within which the experimental conditions prevailing in an experiment can be represented and then the outcome of that experiment can be posed as a constraint in the design and analysis of the next experiment. Description of experimental conditions, in turn, involves both observational and manipulative sentences, and it requires that manipulative phrases (e.g., "having no effect on," "holding Z fixed"), as distinct from observational phrases (e.g., "being independent of," "conditioning on Z"),[4] be given formal notation, semantic interpretation, and axiomatic characterization. It turns out that standard algebras, including the algebra of equations, Boolean algebra, and probability calculus, are all geared to serve observational but not manipulative sentences.

This chapter bases the semantics of manipulative sentences on a set of structural equations that we call a *causal model.* Unlike ordinary algebraic equations, a causal model treats every equation as an independent mathematical object at-

[4]Philosophers, statisticians, and economists have been notoriously sloppy about confusing "holding Z constant" with "conditioning on a given Z" [Pea95a].

tached to one and only one variable. Actions are treated as modalities and are interpreted as the non algebraic operator of replacing equations.

This semantics permits us to develop an axiomatic characterization of manipulative statements of the form "Changing X will not affect Y if we hold Z constant," that we propose as the meaning of causal irrelevance: "X is causally irrelevant to Y in context Z." This axiomatization highlights the differences between causal irrelevance and informational irrelevance, as in "Finding X will not affect our belief in Y, once we know Z." The former shows a closer affinity to graphical representation than the latter. Under the deterministic definition, causal irrelevance complies with all of the axioms of path interception in cyclic graphs except of transitivity. This affinity leads to graphical methods of proving theorems about causal relevance and explains, in part, why graphs are so prevalent in causal talk and causal modeling.

Outside of artificial intelligence, our results have interesting ramifications in the fields of statistics and epidemiology where, thus far the only accepted formalization of causation has been Rubin's framework of counterfactuals [Rub74, Rob86b], which is a rather cumbersome language for expressing causal knowledge. Graphical and structural equation models, popular as they are in econometrics and the social sciences, are viewed with suspicion by statisticians because the causal interpretation of these models has not been adequately formalized [Fre87, Wer92].

Our translation of counterfactuals into statements about structural equation models (Definition 6) generalizes and unifies the structural and counterfactual approaches, and clarifies their conceptual and mathematical bases. The soundness of effectiveness and composition - the only properties of counterfactuals used

by Rubin—ensures that every theorem in Rubin's framework is also a theorem in structural equation models. The completeness of effectiveness and composition in recursive models guarantees that the structural interpretation of counterfactuals introduces no extraneous properties beyond those embodied in Rubin's framework. Most significant, this unification permits investigators to express causal knowledge in the intuitively appealing language of causal graphs, use the graphs as inferential machinery, and be assured of the validity of the results.

$V = \{X, W_1, W_2, W_3, W_4, Y\}$ binary

$U = \{U_1, U_2\}$ binary

$x = u_1$

$w_1 = \neg x \;\&\; \neg u_2$

$w_2 = x \;\&\; \neg u_2$

$w_3 = \neg x \;\&\; u_2$

$w_4 = x \;\&\; u_2$

$y =$

$(w_3 \;\&\; \neg w_1 \;\&\; \neg w_2) \vee$

$(w_4 \;\&\; \neg w_1 \;\&\; \neg w_2)$

$P(u_1) = P(u_2) = 0.5$

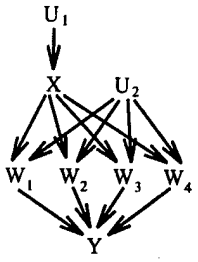


Figure 4.7: Transitivity fails, even when a variable is more completely controlled by its parents

CHAPTER 5

Identifying Causal Effects

5.1 Introduction

This chapter addresses one of the applications of causal models: determining the causal effect of a variable or set of variables on another variable or set of variables. As an introduction, let us consider a possible problem that we would like to solve. Assume we need to replace an expert operating a complex production plant. Before we take charge, we are given a blueprint of the plant together with an explanation of the functions of the various dials and knobs, and we are able to observe the expert in action over a long period of time. During this period, we record which dials the expert consults prior to taking actions. Moreover, although we understand the function of those dials, we cannot always observe the actual reading on each them. The data we are able to collect during the observation period include the actions taken by the agent, the readings of some of the dials, and the outcome of various performance indicators. Our problem is to predict, on the basis of the data collected, the effect of a given action on the performance of the plant.

The problem of learning from the actions of other agents is that one is never sure whether an observed response is due to the agent's action or due to events that triggered that action and simultaneously caused the response. Such events

are called *confounders,* and they present a major problem in the analysis of observational studies in the social and health sciences. For example, we cannot be sure whether it was the prescribed drug, or the some prior condition, which the doctor tried to treat by prescribing the drug, that caused the patient to vomit. Similarly, we cannot tell whether a recession was caused by higher taxes or by the economic indicators which government experts consulted before raising taxes.

The standard technique for dealing with confounders is to adjust for possible variations in those environmental factors which might trigger the actions. This mounts to conditioning the observed distribution on the various levels of those factors, evaluating the action in each level separately, and then taking the (weighted) average over those levels. However, in problems like those described above, some of the confounding factors are unobservable; hence, they cannot be conditioned on.

The techniques developed in this chapter will enable us to recognize, by graphical means, whether a given action can be evaluated from joint distributions on observed quantities and, if so, to decide which quantities should be measured and how to adjust for them. Technically speaking, the task parallels the identification of recursive structural equations in the presence of unmeasured variables. However, whereas traditional theories of identifiability deal exclusively with estimating linear coefficients in parametric equations, the identifiability problem solved in this chapter is nonparametric; no assumptions are made regarding eihter the functional forms of the structural equations or the distributions of the errors.[1]

[1]Naturally, nonparametric identifiability is not concerned with values of numerical parameters but with the ultimate purpose to which parameters are being put in structural models, namely, the analysis of actions and causal effects.

5.2 Identifiability in Econometrics

Determining the causal effect of one variable on another variable has been extensively explored in the econometric literature, where it is called the *identification problem* [Fis66, KR51]. The economic structural equation model consists of M equations in N variables, with M random disturbances. These equations are assumed to be linear in both the variables and the disturbances. Thus, the model can be summed up by the matrix equation

$$A * X = U \tag{5.1}$$

where X is a $1 \times N$ matrix of the observed variables, A is a $M \times N$ matrix of linear coefficients, and U is a $1 \times M$ matrix of random disturbances (which are all zero when considering nonstochastic cases).

The object of the identification problem is to determine the value of the matrix A. In some instances, all of the values in matrix A can be determined; and in other cases, only certain coefficients can be determined. In the econometric literature, when a coefficient can be determined by the data, it is said to be *identifiable*. In the identification problem, we want to find not only a matrix that is observationally equivalent but the actual matrix that determines the interactions of the variables X.

The value of identifying a system of equations is obvious. As soon as we know the equations that model a system, we know everything about how the system behaves. We can determine the value of any set of variables under all possible interventions on other variables. We can consider how changes to the model will affect various variables. Thus, we can use such a system to inform policy decisions. Even if we can only identify some of the coefficients, in many cases we

can still determine the direct or total effect of one variable on another.

There are, however, many limitations to this approach. One of the greatest is the linearity assumption. We often cannot assume that all of the interactions are linear, yet if we allow each variable to be an arbitrary function of the other variables in the system, then there is no way, in general, to completely determine the values of the functions, as we can in the linear case. All is not lost, however. Let us consider why we wanted to find the values of the functions in the first place. We need the actual equations, instead of observationally equivalent ones, only when we want to predict how the system would react to interventions that are outside the measured values of the variables. We will show how we can obtain information about how the system responds to interventions, even without completely specifying the equations in the model.

5.3 Identification in Causal Models

In the language of causal models, the problem addressed in this chapter is the evaluation of the effects of a concurrent action $do(X = x)$, where X is some subset of variables from V, on a subset of variables Y. We will be examining such actions for the case where the causal model is not fully specified. We are given the topology of the causal model but not the actual functions that relate the variables to each other. Numerical probabilities are given for only the variables which are deemed "observable," while the variables deemed "unobservable" serve only to specify possible connections among observed quantities, and are not given numerical probabilities.

Pearl [Pea94] has reviewed the use of causal models in this fashion and proposed a calculus for deriving probabilistic assessments of the effects of actions

in the presence of unmeasured variables. This calculus can be used to check or search for a proof that the effect of one variable on another is *identifiable*, namely, that it is possible from data involving only observed variables, to obtain a consistent estimate of the probability of Y under the condition that X is set to x by external intervention. This chapter systematizes the search for such a proof by providing a polynomial-time graph-based method for determining whether the effect of one variable on another is identifiable.[2]

If identifiability is confirmed, the method generates closed-form expressions for the distribution of the outcome variable Y under the external manipulation of the control variable X. The derived expression, denoted $P(y|do(x))$, invokes only measured probabilities as obtained, for example, by recording the past performance of other acting agents. Although the actions of those agents may have been triggered by factors unseen by the analyst, the impact of X on Y can still be predicted using observed variables only. If Y stands for a goal variable, then the probability of reaching the goal through each action $do(X = x)$ can be determined from such partial observations.

5.4 Notation and Definitions

We now provide some notation and definitions that will be required in the rest of the chapter.

Definition 23 (identifiability) *The causal effect of X on Y is said to be* identifiable *if the quantity $P(y|\hat{x})$ can be computed uniquely from any positive distribution of the observed variables, that is, if for every pair of models M_1 and M_2 such that*

[2] An extension of our analysis to the case of multiple actions (sequential or concurrent) is reported in [PR95].

$P_{M_1}(\mathbf{v}) = P_{M_2}(\mathbf{v}) > 0$, *we have* $P_{M_1}(y|\hat{x}) = P_{M_2}(y|\hat{x})$

Identifiability means that $P(y|\hat{x})$ can be estimated consistently from an arbitrarily large sample randomly drawn from the distribution of the observed variables.

Definition 24 (back-door path) *A path from X to Y in a graph G is said to be a* back-door path *if it contains an arrow into X.*

The probabilistic analysis of causal models becomes particularly simple when two conditions are satisfied:

1. The model is recursive, that is, there exists an ordering of the variables $V = \{X_1, \ldots, X_n\}$ such that each X_i is a function of a subset $\mathbf{pa}_i$ of its predecessors, denoted

$$X_i = f_i(\mathbf{pa}_i, U_i), \qquad \mathbf{pa}_i \subseteq \{X_1, \ldots, X_{i-1}\} \tag{5.2}$$

2. The disturbances $U_1, \ldots, U_n$ are mutually independent, $U_i \perp U_j$, which also implies (from the exogeneity of the U_i's) that

$$U_i \perp \{X_1, \ldots, X_{i-1}\} \tag{5.3}$$

These two conditions, also called the *Markovian assumptions*, are the basis of Bayesian networks [Pea88], and they enable us to compute causal effects directly from the conditional probabilities $P(x_i|\mathbf{pa}_i)$ without having to specify either the functional form of the functions f_i or the distributions $P(u_i)$ of the disturbances. This is seen immediately from the following observations.

The distribution induced by any Markovian model M is given by the product

$$P_M(x_1, \ldots, x_n) = \prod_i P(x_i|\mathbf{pa}_i) \tag{5.4}$$

where $\mathbf{pa}_i$ are the direct predecessors (called *parents*) of X_i in the diagram. The distribution induced by the the submodel $M_{x_j'}$, which represents the action $do(X_j = x_j')$, is also Markovian and, hence, also induces a product-like distribution:

$$P_{M_{x_j'}}(x_1, \ldots, x_n)$$

$$= \begin{cases} \prod_{i \neq j} P(x_i|\mathbf{pa}_i) = \frac{P(x_1, \ldots, x_n)}{P(x_j|\mathbf{pa}_j)} & \text{if } x_j = x_j' \\ 0 & \text{if } x_j \neq x_j' \end{cases} \tag{5.5}$$

where the partial product reflects the surgical removal of

$$X_j = f_j(\mathbf{pa}_j, U_j)$$

from the model of Eq. (5.2).

5.5 Action Calculus

The identifiability of causal effects demonstrated in Section 5.3 relies critically on the the two Markovian assumptions given by Eqs. (5.2) and (5.3). If a variable that has two descendants in the graph is unobserved, the disturbances in the equations for those two descendants are no longer independent, the Markovian assumtion given by Eq. (5.2) is violated, and identifiability may be destroyed. This can be seen easily from Eq. (5.5): if any parent of the manipulated variable X_j is unobserved, one cannot estimate the conditional probability $P(x_j|\mathbf{pa}_j)$, and

the effect of the action $do(X_j = x_j)$ may not be predictable from the observed distribution $P(x_1, \ldots, x_n)$. Fortunately, certain causal effects are identifiable even in situations where members of $\mathbf{pa}_j$ are be unobservable, and these situations can be recognized through the action calculus introduced in [Pea94].

Let X, Y, and Z be arbitrary disjoint sets of nodes in a Directed Acyclic Graph (DAG) G. We denote by $G_{\overline{X}}$ the graph obtained by deleting from G all arrows pointing to nodes in X. Likewise, we denote by $G_{\underline{X}}$ the graph obtained by deleting from G all arrows emerging from nodes in X. To represent the deletion of both incoming and outgoing arrows, we use the notation $G_{\overline{X}\underline{Z}}$. Finally, the expression $P(y|\hat{x}, z) \triangleq P(y, z|\hat{x})/P(z|\hat{x})$ stands for the probability of $Y = y$ given that $Z = z$ is observed and X is held constant at x.

If G is a DAG, let $G_{\overline{X}}$ stand for the subgraph of G with all the arcs incident to variables in X removed, and $G_{\underline{X}}$ stand for the subgraph of G with all the arcs emanating from X removed. Likwise, let $(Y \perp\!\!\!\perp X|Z)_G$ stand for Y being d-separated from X by Z in the graph G [Pea88].

The following theorem states the three basic inference rules used in the chapter.

Theorem 11 *Let G be the directed acyclic graph associated with a causal model, and let $P(\cdot)$ stand for the probability distribution induced by that model. For any disjoint subsets of variables X, Y, Z, and W we have:*

Rule 1 (Insertion/deletion of observations)

$$P(y|\hat{x}, z, w) = P(y|\hat{x}, w) \quad \textit{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}} \tag{5.6}$$

Rule 2 (Action/observation exchange)

$$P(y|\hat{x},\hat{z},w) = P(y|\hat{x},z,w) \quad if \quad (Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}\underline{Z}}} \tag{5.7}$$

Rule 3 (Insertion/deletion of actions)

$$P(y|\hat{x},\hat{z},w) = P(y|\hat{x},w) \quad if \quad (Y \perp\!\!\!\perp Z|X,\ W)_{G_{\overline{X},\ \overline{Z(W)}}} \tag{5.8}$$

where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$.

Each of these inference rules follows from the basic interpretation of the "$\hat{x}$" operator as the replacement of the causal mechanism that connects X to its pre-action parents by a new mechanism $X = x$ introduced by the intervening force. The result is a submodel characterized by the subgraph $G_{\overline{X}}$ (named manipulated graph in [SGS93]) that supports all three rules.

Rule 1 reaffirms d-separation as a valid test for conditional independence in the distribution resulting from the intervention $set(X = x)$, hence the graph $G_{\overline{X}}$. This rule follows from the fact that deleting equations from the system does not introduce any dependencies among the remaining disturbance terms.

Rule 2 provides a condition for an external intervention $set(Z = z)$ to have the same effect on Y as the passive observation $Z = z$. The condition amounts to $\{X \cup W\}$ blocking all back-door paths from Z to Y (in $G_{\overline{X}}$), since $G_{\overline{X}\underline{Z}}$ retains all (and only) such paths.

Rule 3 provides conditions for introducing (or deleting) an external intervention $set(Z = z)$ without affecting the probability of $Y = y$. The validity of this rule stems, again, from simulating the intervention $set(Z = z)$ by the deletion of all equations corresponding to the variables in Z (hence the graph $G_{\overline{XZ}}$).

Corollary 2 *A causal effect $q = P(y_1, ..., y_k|\hat{x}_1, ..., \hat{x}_m)$ is identifiable in a model characterized by a graph G if there exists a finite sequence of transformations, each conforming to one of the inference rules in Theorem 11, which reduces q to a standard (i.e., hat-free) probability expression involving observed quantities.*

Although Theorem 11 and Corollary 2 require that the Markovian assumtions hold, they can be applied to recursive non-Markovian models, because such models become Markovian if we consider the unobserved variables as part of the analysis and represent these variables as nodes in the graph.

5.6 A Graphical Criterion for Testing Identifiability

To avoid excessive notation, for the rest of this chapter, we will consistently refer to queries $P(y|\hat{x})$ that satisfy Corollary 2 as "identifiable," with the understanding that Corollary 2 represents a sufficient but not (yet) necessary condition for semantical identifiability, given in Definition 23. The two notions would be equivalent if the rules in Theorem 11 were complete.

Theorem 12 *A necessary and sufficient condition for the identifiability of $P(y|\hat{x})$ in a graph G is that G satisfies one of the following four conditions:*

1. There is no back-door path from X to Y in G, that is, $(X \underline{\perp\!\!\!\perp} Y)_{G_{\underline{X}}}$.
2. There is no directed path from X to Y in G.
3. There exists a set of nodes B that blocks all back-door paths from X to Y such that $P(b|\hat{x})$ is identifiable. A special case of this condition occurs when B consists entirely of nondescendants of X, in which case $P(b|\hat{x})$ reduces immediately to $P(b)$.

4. There exists set of nodes Z_1 and Z_2 such that

   (i) Z_1 blocks every directed path from X to Y,

   i.e., $(Y \perp\!\!\!\perp X|Z_1)_{G_{\overline{Z_1}\,\overline{X}}}$

   (ii) Z_2 blocks all back-door paths between Z_1 and Y,

   i.e., $(Y \perp\!\!\!\perp Z_1|Z_2)_{G_{\overline{X}\underline{Z_1}}}$,

   (iii) Z_2 blocks all back-door paths between X and Z_1,

   i.e., $(X \perp\!\!\!\perp Z_1|Z_2)_{G_{\underline{X}}}$,

   (iv) Z_2 does not conduct any back-door paths from X to Y,

   i.e., $(X \perp\!\!\!\perp Y|Z_1, Z_2)_{G_{\overline{Z_1}\,\overline{X(Z_2)}}}$. This condition holds if the conditions (i)–(iii) above are met and no member of Z_2 is a descendant of X.

A special case of Condition 4 occurs when $Z_2 = \emptyset$ and there is no back door path from X to Z_1 or from Z_1 to Y.

Proof (of Theorem 12):

We prove the sufficiency of Conditions (1)–(4) above, then turn to proving their necessity.

Condition 1: If there is no directed path from X to Y in G, then $(Y \perp\!\!\!\perp X)_{G_{\overline{X}}}$. So, by Rule 3, $P(y|\hat{x}) = P(y)$, and the query is identifiable.

Condition 2: This condition follows directly from Rule 1. If $(Y \perp\!\!\!\perp X)_{G_{\underline{X}}}$, then we can immediately change $P(y|\hat{x})$ to $P(y|x)$, so the query is identifiable.

Condition 3: If there is a set of nodes B that blocks all back-door paths from X to Y, then we can rewrite $P(y|\hat{x})$ as $\sum_b P(y|\hat{x}, b)P(b|\hat{x})$. Since B blocks all back-door paths from X to Y, it must be the case that $(Y \perp\!\!\!\perp X|B)_{G_{\underline{X}}}$, and thus,

by Rule 2, we can rewrite $P(y|\hat{x}, b)$ as $P(y|x, b)$. If the query $(b|\hat{x})$ is identifiable, then the original query must also be identifiable. See examples in Figure 5.1.

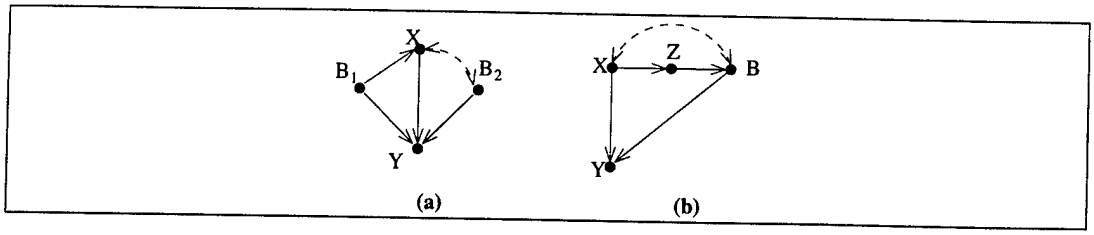


Figure 5.1: Illustrating Condition 3 of Theorem 12. In **a**, the set $\{B_1, B_2\}$ blocks all back-door paths from X to Y and $P(b_1, b_2|\hat{x}) = P(b_1, b_2)$. In **b**, the node B blocks all back-door paths from X to Y, and $P(b|\hat{x})$ is identifiable using Condition 4.

Condition 4: If there is a set of nodes Z_1 that block all directed paths from X to Y and a set of nodes Z_2 that block all back-door paths between Y and Z_1 in $G_{\overline{X}}$, that we expand $P(y|\hat{x}) = \sum_{z_1,z_2} P(y|\hat{x}, z_1, z_2)P(z_1, z_2|\hat{x})$. we can rewrite $P(y|\hat{x}, z_1, z_2)$ as $P(y|\hat{x}, \hat{z}_1, z_2)$ using Rule 2, since all back-door paths between Z_1 and Y are blocked by Z_2 in $G_{\overline{X}}$. We can reduce $P(y|\hat{x}, \hat{z}_1, z_2)$ to $P(y|\hat{z}_1, z_2)$ using Rule 3, since $(Y \perp\!\!\!\perp X|Z_1, Z_2)_{G_{\overline{Z_1}\,\overline{X(Z_2)}}}$. We can rewrite $P(y|\hat{z}_1, z_2)$ as $P(y|z_1, z_2)$ if $(Y \perp\!\!\!\perp Z_1|Z_2)_{G_{\underline{Z_1}}}$. The only way that this independence cannot hold is if there is a path from Y to Z_1 through X, since $(Y \perp\!\!\!\perp Z_1|Z_2)_{G_{\overline{X}\underline{Z_1}}}$. However, we can block this path by conditioning and summing over X, to get $\sum_{x'} P(y|\hat{z}_1, z_2, x')P(x'|\hat{z}_1, z_2)$. Now we can rewrite $P(y|\hat{z}_1, z_2, x')$ as $P(y|z_1, z_2, x')$ using Rule 2. $P(x'|\hat{z}_1, z_2)$ can be rewritten as $P(x'|z_2)$ using Rule 3, since Z_1 is a child of X and the graph is acyclic. So, the query can be rewritten as $\sum_{z_1,z_2} \sum_{x'} P(y|z_1, z_2, x')P(x'|z_2)P(z_1, z_2|\hat{x})$. $P(z_1, z_2|\hat{x}) = P(z_2|\hat{x})P(z_1|\hat{x}, z_2)$. Since Z_2 consists of non-descendants of X, we can rewrite $P(z_2|\hat{x})$ as $P(z_2)$ us-

ing Rule 3. Since Z_2 blocks all back-door paths from X to Z_1, we can rewrite $P(z_1|\hat{x}, z_2)$ as $P(z_1|x, z_2)$ using Rule 2. The entire query can thus be rewritten as $\sum_{z_1,z_2} \sum_{x'} P(y|z_1, z_2, x')P(x'|z_2)P(z_1|x, z_2)P(z_2)$. See examples in Figure 5.2

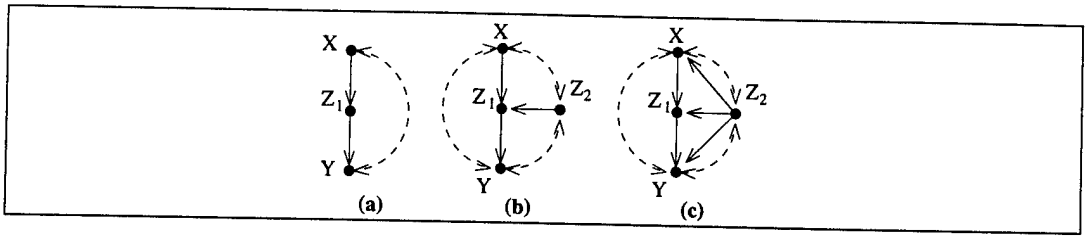


Figure 5.2: Illustrating Condition 4 of Theorem 12. **(a)**, Z_1 blocks all directed paths from X to Y, and the empty set blocks all back-door paths from Z_1 to Y in $G_{\overline{X}}$ and all back-door paths from X to Z_1 in G; **(b,c)** Z_1 blocks all directed paths from X to Y, and Z_2 blocks all back-door paths from Z_1 to Y in $G_{\overline{X}}$ and all back-door paths from X to Z_1 in G

We now prove that the conditions of Theorem 12 are necessary. This may be shown by contradiction

Proof Sketch: We will assume that there exists a query $P(y|\hat{x})$ and a graph G such that (1) None of the conditions of Theorem 12 holds, and (2) there exists a finite sequence of applications of inference rules which removes all hats from the variables in the query. We will show that these two assumptions lead to a contradiction; hence, if all four conditions of Theorem 12 fail, there must not be a finite sequence of inference rules that reduces the query to a hat-free expression.

Proof Outline:

I $(Y \perp\!\!\!\perp X|Z, W)_{G_{\overline{Z}\underline{X}}}$, so Rule 2 can be applied to remove the hat from X.

A There is a directed path from Z to Y

1 Cannot add $\hat{z}$ using Rule 3

2 Cannot add $\hat{z}$ using Rule 2

B There is a directed path from Z to X

1 Cannot remove $\hat{z}$ using Rule 2

2 Cannot remove $\hat{z}$ using Rule 3

II $(Y \perp\!\!\!\perp X|Z,W)_{G_{\overline{Z}\,\overline{X(W)}}}$, so Rule 3 can be applied to remove $\hat{x}$

A Cannot add $\hat{z}$ using Rule 3

B Cannot add $\hat{z}$ using Rule 2

Assume that there exists a query $P(y|\hat{x})$ and a graph G such that none of the conditions of Theorem 12 holds, but the query is still identifiable. Since $P(y|\hat{x})$ is identifiable, there must be some finite sequence of inference rules that removes the hat from X. This means that there must be some (possibly empty) set of variables Z and W such that either $(Y \perp\!\!\!\perp X|Z,W)_{G_{\overline{Z}\underline{X}}}$, so we can reduce $P(y|\hat{x},\hat{z},w)$ to $P(y|x,\hat{z},w)$ via Rule 2, or $(Y \perp\!\!\!\perp X|Z,W)_{G_{\overline{Z}\,\overline{X(W)}}}$, so we can reduce $P(y|\hat{x},\hat{z},w)$ to $P(y|\hat{z},w)$ using Rule 3. We will look at each of these cases in turn.

Case I: Consider $(Y \perp\!\!\!\perp X|Z,W)_{G_{\overline{Z}\underline{X}}}$. By assumption, $P(y|\hat{x})$ is identifiable, and the hat is removed from X by an application of Rule 2. This implies a series of rule applications to $P(y|\hat{x})$ which results in $P(y|\hat{x},\hat{z},w)$ such that $(Y \perp\!\!\!\perp X|Z,W)_{G_{\overline{Z}\underline{X}}}$. We will look at the restrictions on imposed Z and W by both the failure of the conditions of Theorem 12 to hold and the assumption that $P(y|\hat{x})$ can be transformed to $P(y|\hat{x},\hat{z},w)$ by a series of rule applications. We will also make the assumption that Z and W are minimal. If they are not, then there exist minimal Z' and W', in which superfluous nodes are removed, that

would work. Thus, proving that no minimal Z' and W' exist implies that no Z and W exist.

If $(Y \perp\!\!\!\perp X|Z,W)_{G_{\underline{X}}}$, then a blockable back-door path would exist, and Condition 3 of Theorem 12 would have held. We also know $(Y \perp\!\!\!\perp X|Z,W)_{G_{\overline{Z}\underline{X}}}$, by assumption. These two independence assertions imply that Z conducts a back-door path that is not blocked by W. That is, there is a back-door path between X and Y that has a head-to-head junction in Z. Each element of Z must also block a back-door path from X to Y, since Z is minimal. This implies that there is a directed path from Z to X or from Z to Y (Figure 5.3).

Proof that there is a directed path from Z to X or from Z to Y: Since we know that Z must block a back-door path from X to Y, there must be a path from Z to X or from Z to Y that starts in an arrow that is incident away from Z. All of the head-to-head junctions along this path must either be in W or have descendants in W. If there are no such head-to-head junction paths, then there is a directed path from Z to X or from Z to Y. If there is a head-to-head junction, then consider the W that unblocks this junction. This W must itself block a back-door path from X to Y, so, there must be a path from W to either X or Y that starts with an arc incident away from W. This path is either a directed path from W to X or from W to Y, or has a head-to-head junction that is also a member of W or an ancestor of a member of W. Since the graph is acyclic, there must eventually be a W that has a directed path to X or Y that is also a descendant of Z. Thus there is a directed path from Z to either X or Y.

We now look at the cases of Case I.

Case IA: A directed path exists from Z to Y. By our assumption, there must be a sequence of rules that transforms $P(y|\hat{x})$ to $P(y|\hat{x},\hat{z},w)$. There are

two ways of adding $\hat{z}$ to $P(y|\hat{x})$—by using Rule 3, or by first conditioning on Z and then adding the hat to it by using Rule 2.

Case IA1: First we look at using Rule 3. If there is a directed path from Z to Y (Figure 5.3a), then $(Y \underline{\not\perp} Z|X)_{G_{\overline{X}\,\overline{Z}}}$. No element of W can block this path from Y to Z, since that would require W to be a descendant of Z, and $(Y \underline{\not\perp} Z|X, W)_{G_{\overline{X}}}$. So Rule 3 cannot be invoked to add $\hat{z}$ to $P(y|\hat{x})$.

Case IA2: We need to condition on Z and then add the hat to it using Rule 2. In order for us to add the hat to Z using Rule 2, there needs to be a W' such that $(Y \underline{\perp\!\!\!\perp} Z|W', X)_{G_{\overline{X}\underline{Z}}}$. Above, we proved that given our assumptions, there must be an unblocked path from Y to X that has a head-to-head junction at Z and that no member of W blocks this path, so $W' \not\subseteq W$. If we condition on a W' that allows us to add the hat to Z, we must then remove this W' to obtain $P(y|\hat{x}, \hat{z}, w)$ so that we can remove the hat from X. However, we are not able to remove this W'. We cannot remove W' using Rule 1, since $(Y \underline{\not\perp} W'|X, Z, W)_{G_{\overline{X}\,\overline{Z}}}$, and if we add some W'' that d-separates Y from W', then we would not be able to remove W''. Thus, we cannot add $\hat{z}$ to $P(y|\hat{x})$ by first conditioning on Z and then adding the hat to it by using Rule 2 if there is a directed path from Z to Y.

Case IB: A directed path exists from Z to X. If there is a directed path from Z to X (Figure 5.3b), we can assume that we can add $\hat{z}$ to $P(y|\hat{x})$ to get $P(y|\hat{x}, \hat{z})$, condition on W to get $P(y|\hat{x}, \hat{z}, w)$, and then use Rule 2 to remove the hat from X.

Now we will prove that there is no way to remove $\hat{z}$ from the expression $P(y|x, \hat{z}, w)$. Since there is a back-door path from X to Y that has a head-to-head junction at Z, there must be a back-door path from Z to Y.

Case IB1: If we could remove the hat from Z using Rule 2, then we could block the back-door path from Z to Y and, hence, block the back-door path from X to Y, and Condition 3 would have held.

Case IB2: If we could remove $\hat{z}$ directly using Rule 3, then there would have to be some set of nodes that blocked the directed path from Z to X, and both $(Y \perp\!\!\!\perp X|Z,W)_{G_{\overline{Z}\underline{X}}}$ and $(Y \not\perp\!\!\!\perp X|Z,W)_{G_{\underline{X}}}$ would not be true.

Thus, we cannot remove all the hats from the expression by using Rule 2 to remove the hat from X.

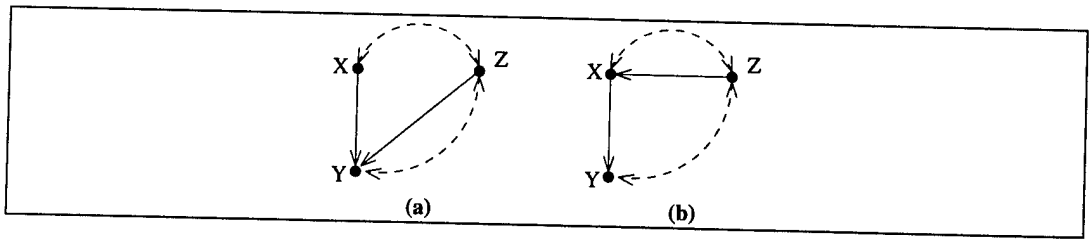


Figure 5.3: Using Rule 2 to remove the hat from X when the criterion fails: since Z is necessary, there must be a directed path from **(a)** Z to Y or **(b)** Z to X

Case II: Now consider $(Y \perp\!\!\!\perp X|Z,W)_{G_{\overline{Z}\,\overline{X(W)}}}$. We will try to find a set of rule applications that transforms $P(y|\hat{x})$ into $P(y|\hat{x},\hat{z},w)$ when none of the conditions of Theorem 12 holds. Z must block all directed paths from X to Y. If it did not, then W would have to block a directed path, which would make W a descendant of X, so then $X(W) = \emptyset$, and thus $(Y \perp\!\!\!\perp X|Z,W)_{G_{\overline{Z}}}$; however, we proved above that this could not happen if any of the conditions of Theorem 12 holds. There are two ways to of adding $\hat{z}$ to $P(y|\hat{x})$—by using Rule 3 directly, or by conditioning on Z and then adding the hat to it using Rule 2. We will look at each of these in turn.

Case IIA: First, we will try to add $\hat{z}$ directly by using Rule 3. To do this, there must be some W such that $(Y \perp\!\!\!\perp Z|W,X)_{G_{\overline{X}\,\overline{Z(W)}}}$. Since there is a directed path from Z to Y, W must be a descendant of Z and thus $(Y \perp\!\!\!\perp Z|W,X)_{G_{\overline{X}}}$ So, W blocks all back-door paths between Z and Y in $G_{\overline{X}}$. Once $\hat{x}$ has been removed from $P(y|\hat{x},\hat{z},w)$ to obtain $P(y|\hat{z},w)$, we need to remove $\hat{z}$, or remove the hat from Z. We cannot remove the hat from Z directly by using Rule 3, since $Z(W) = \emptyset$ and thus $(Y \not\perp\!\!\!\perp Z|W,X)_{G_{\overline{Z(W)}}}$, and there is a back door path from Z to Y through X. If we could remove the hat from Z by using Rule 2, then Condition 4 would hold. So, we cannot add $\hat{z}$ directly by using Rule 3 if any of the conditions of Theorem 12 holds.

Case IIB: Next, we will try to condition on Z and then add the hat to it by using Rule 2. However, for this to be possible, there would have to be a W that blocks all back door-paths between X and Z, and between Z and Y—and then Condition 4 would hold.

Thus, if none of the conditions of Theorem 12 hold, the query must not be identifiable. □

Remark: The criterion in Theorem 12 is complete only if the inference rules themselves are complete. We will look at each of the three rules in Theorem 11 and show that the graphical conditions that license each are the tightest possible.

$$(Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}}} \quad \textit{if} \quad P(y|\hat{x},z,w) = P(y|\hat{x},z)$$

Since the d-separation condition is valid for any recursive model, including the submodel represented by $G_{\overline{X}}$, the conditional independence $P(y|\hat{x},z,w) = P(y|\hat{x},z)$ implies $(Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}}}$.

$$(Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}\underline{Z}}} \quad \textit{if} \quad P(y|\hat{x},\hat{z},w) = P(y|\hat{x},z,w)$$

Consider the augmented diagram G' that has the intervention arcs $F_Z \rightarrow Z$ added. $P(y|\hat{x},\hat{z},w) = P(y|\hat{x},z,w)$ implies that $(Y \perp\!\!\!\perp F_Z|X,Z,W)_{G'_{\overline{X}}}$. If there is a path from Y to Z that is unblocked by $\{X,W\}$ in $G_{\overline{X}}$, this path must not end in an arrow incident to Z, if it did $(Y \perp\!\!\!\perp F_Z|X,Z,W)_{G'_{\overline{X}}}$ would not hold. Since every path from Y to Z that is not blocked by $\{X,W\}$ in $G_{\overline{X}}$ must pass through an arrow leaving Z, $(Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}\underline{Z}}}$.

$$(Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}\,\overline{Z(W)}}} \quad \textit{if} \quad P(y|\hat{x},\hat{z},w) = P(y|\hat{x},w)$$

Again consider G' with intervention arcs $F_Z \rightarrow Z$ added. $P(y|\hat{x},\hat{z},w) = P(y|\hat{x},w)$ implies that $(Y \perp\!\!\!\perp F_Z|X,W)_{G'_{\overline{X}}}$. Hence, any path from Z to Y that is not blocked by $\{X,W\}$ in $G_{\overline{X}}$ must end in an arrow pointing to Z; otherwise, $(Y \perp\!\!\!\perp F_Z|X,W)_{G'_{\overline{X}}}$ would not hold. In addition, if there is a path from some Z' of Z to Y that does end in an arrow pointing to Z', then W must not be a descendant of Z'; otherwise, $(Y \perp\!\!\!\perp F_Z|X,W)_{G'_{\overline{X}}}$ would not hold. Thus, the only paths from Y to Z must end in an arrow pointing at Z and in some member of $Z(W)$. Thus, $(Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}\,\overline{Z(W)}}}$.

Although these rules are as tight as possible, some strange exchange of hatted and hatless variables that is not reachable by successive applications of Rules 1–3 might still be licensed by some graph. Thus, it is possible that the three inference rules (and hence the Theorem 12) are not complete.

5.7 Remarks on Efficiency

In implementing Theorem 12 as a systematic method for determining identifiability, Conditions 3 and 4 would seem to require exhaustive search. To prove that Condition 3 does not hold, for instance, we need to prove that no blocking set B can exist. Fortunately, the following theorems allow us to significantly prune the search space, so as to render the test tractable.

Theorem 13 *If for one minimal set B_i, $P(b_i|\hat{x})$ is identifiable, then for any other minimal set B_j, $P(b_j|\hat{x})$ is identifiable.*

Theorem 13 allows us to test Condition 3 with a single minimal blocking set B. If B meets the requirements for Condition 3, then the query is identifiable; otherwise, Condition 3 cannot be satisfied. In proving this theorem, we use the following lemma.

Lemma 2 *If the query $P(y|\hat{x})$ is identifiable, and a set of nodes Z lies on a directed path from X to Y, then the query $P(z|\hat{x})$ is identifiable.*

Theorem 14 *Let Y_1 and Y_2 be two subsets of nodes such that either no nodes Y_1 are descendants of X, or all nodes Y_1 and Y_2 are descendants of X and all nodes Y_1 are nondescendants of Y_2. A reducing sequence for $P(y_1, y_2|\hat{x})$ exists (per Corollary 1) iff there are reducing sequences for both $P(y_1|\hat{x})$ and $P(y_2|\hat{x}, y_1)$.*

$P(y_1, y_2|\hat{x})$ may possibly pass the test in Theorem 12 if we apply the procedure to both $P(y_2|\hat{x}, y_1)$ and $P(y_1|\hat{x})$, but if we try to apply the test to $P(y_1|\hat{x}, y_2)$, we will not find a reducing sequence of rules. Figure 5.4 shows just such an example. Theorem 14 guarantees that, if there is a reducing sequence for $P(y_1, y_2|\hat{x})$, then

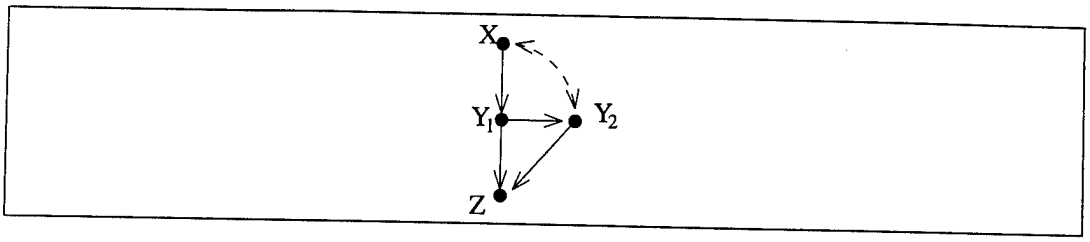


Figure 5.4: Theorem 12 ensures a reducing sequence for $P(y_2|\hat{x}, y_1)$ and $P(y_1|\hat{x})$, although none exists for $P(y_1|\hat{x}, y_2)$

we should always be able to find such a sequence for both $P(y_1|\hat{x})$ and $P(y_2|\hat{x}, y_1)$ by proper choice of Y_1.

Theorem 15 *If there exists a set Z_1 that meets all of the requirements for Z_1 in Condition 4, then the set consisting of the children of X intersected with the ancestors of Y will also meet all of the requirements for Z_1 in Condition 4.*

Theorem 15 removes the need to search for Z_1 in Condition 4 of Theorem 12.

We now provide proofs for Theorems 13–15

Proof (of Theorem 13):

(By contradiction.)

Assume that there is a minimal set B such that $(Y \perp\!\!\!\perp X|B)_{G_{\underline{X}}}$ and the query $P(b|\hat{x})$ is identifiable. Assume that there is another minimal set K such that $(K \perp\!\!\!\perp X|B)_{G_{\underline{X}}}$ and the query $P(k|\hat{x})$ is not identifiable.

Consider all (undirected) paths from X to Y in $G_{\underline{X}}$. Every element of B and K must lie along one of these paths, since the sets are minimal. In addition, at least one member of K must be a descendant of X, otherwise $P(k|\hat{x})$ would be identifiable. In fact, any member of K that is a descendant of X needs to lie on

a directed path from X to Y. To see that this is true, note that if a member K_1 of K is a descendant of X but does not lie on a directed path from X to Y, then there must be a head-to-head junction along the path from K_1 to Y. This path would have to be unblocked by some other member K_2 of K. Since K is minimal, there must be some unblocked path from some descendant of K_2 to Y that K blocks. This implies that there is either a directed path from one of the descendants of K_2 to Y, which would make K_1 an ancestor of Y, or a head-to-head junction on the path from K_2 to Y that is unblocked by some other member K_3 of K. Namely, there is either an infinite series of Ks between K_1 and Y, or else a directed path from K_1 to Y (see Figure 5.5).

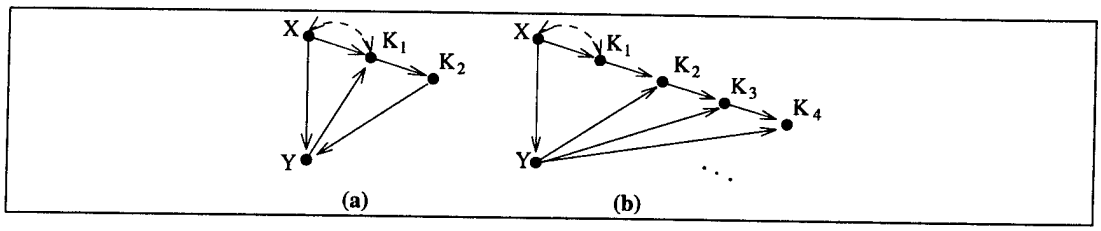


Figure 5.5: If a member of K blocks a back-door path from X to Y and is a descendant of X, then it is also an ancestor of Y

Let K' be the subset of K that lies on a directed path from X to Y, and let $K'' = K \setminus K'$. We know that $P(k|\hat{x}) = P(k'|\hat{x}, k'') * P(k''|\hat{x})$ and that $P(k''|\hat{x}) = P(k'')$. So, $P(k'|\hat{x}, k'')$ must not be identifiable. Since K is minimal, K' must block some back-door path, and that back-door path must also be blocked by some member B' of B. There are two possibilities: either the path that K' blocks has a head-to-head junction that is not unblocked by B or there is some member B' of B which blocks the same back-door path. These two cases are illustrated in Figure 5.6.

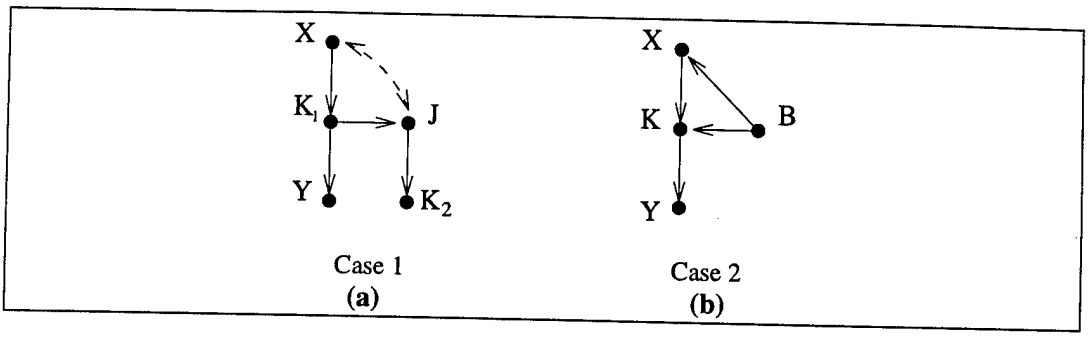


Figure 5.6: Examples of the two cases for K'

Case 1: There is a head-to-head junction that is not unblocked in B, but is unblocked in K. Call this junction J. Since K is minimal, the element of K that unblocks this path (equal to either J or one of J's descendants) must lie on some unblocked path from Y to X in $G_{\underline{X}}$. If this is the case, then there must be an unblocked path through J's descendants that also goes through J, which means there must be some element B' of B that blocks the path between J and X in $G_{\underline{X}}$ (see Figure 5.7). We can condition and sum over this B' to get

$$\begin{aligned} P(k'|\hat{x}, k'') &= \sum_{b'} P(k'|\hat{x}, k'', b') * P(b'|\hat{x}, k'') \\ &= \sum_{b'} P(k'|x, k'', b') * P(b'|\hat{x}, k'') \end{aligned}$$

by using Rule 2. So, the query $P(b'|\hat{x}, k'')$ must not be identifiable. Thus, B' must be a descendant of X, because otherwise $P(b'|\hat{x}, k'') = P(b'|k'')$. So, $P(b'|\hat{x})$ is identifiable, but $P(b'|\hat{x}, k'')$ is not. Therefore, K'' must disallow the blocking of a back-door path from X to B'. As a result, there must be a back-door path from X to B' that has a head-to-head junction, and this junction must have a descendant in K'' but not in B. This is impossible: since K is minimal, the descendant of the head-to-head junction must block a back-door path from X to

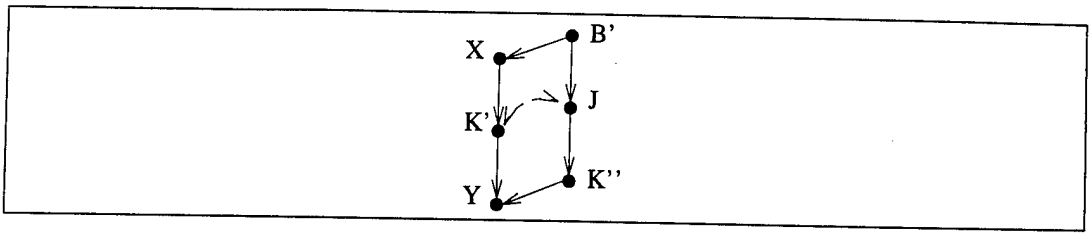


Figure 5.7: There must exist a member B' of B which blocks the back-door path from X to J

Y. B must block that same path, meaning the path from X to B' was unblocked by B as well as by K''.

Case 2: There is a member B' of B that blocks the same back-door path as K'. The path could be blocked by B' bewteen either X and K', or between K' and Y (See Figure 5.8). If the path is blocked by B' between X and K', we have the same contradiction as in Case 1 above. If it is blocked by B' between K' and Y, then B' lies on a directed path from X to Y. From Lemma 2, we know that $P(k'|\hat{x})$ must be identifiable. That means that K'' must disallow either Condition 3 or Condition 4 of the Theorem 12. If it blocks Condition 3, then K'' must conduct a back-door path from X to K'. Namely, some member of K'' is at or is a descendant of a head-to-head junction along a path from K' to X in $G_{\underline{X}}$. Using the same argument as above, since K is minimal, the path blocked by K'' must also be blocked by B, and thus the head-to-head junction must be unblocked by B as well. Any unblockable back-door path from X to K' will also be an unblockable back-door path from X to B', since B' is a direct descendant of K'. However, we know that there cannot be a back-door path from X to B' that is unblockable when we condition on B. Thus there cannot be a back-door path from X to K' that is unblockable when we condition on K''.

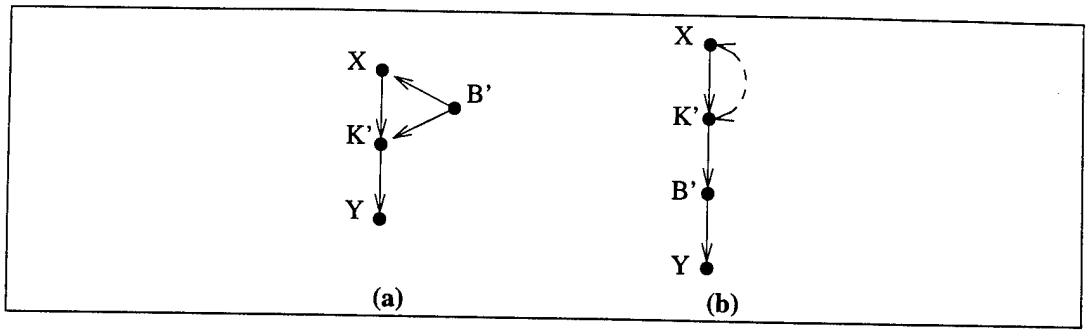


Figure 5.8: B can be between either **(a)** X and K', or **(b)** K' and Y.

If K'' disallows Condition 4, then some other set of nodes R must block every directed path from X to K'. K'' must unblock a back-door path from X to R or from R to K'. As above, if a back-door path from X to R (and thus from X to K') is unblocked by K'', it will also be unblocked by B. So, K'' must unblock a back-door path from R to K'. Since K is minimal, there must be a path from a descendant of R to X in $G_{\underline{X}}$, which implies that there must also be a path from Y to X in $G_{\underline{X}}$ that passes through R and K'. Since the back-door path from X to K' must not be blockable (since a blockable back-door path was invalidated above), B must block the path from Y to K'. But then there would not be a back-door path from R to K' that is blockable when conditioning on B but unblockable when conditioning on K''.

So, if any minimal set B blocks all back-door paths from X to Y and the query $P(b|\hat{x})$ is identifiable, then if any other minimal set K blocks all back-door paths from X to Y, $P(k|\hat{x})$ must also be identifiable. □

Proof (of Lemma 2):

If the query $P(y|\hat{x})$ is identifiable, one of the four conditions of Theorem 12 must have been satisfied. We Look at each in turn.

Condition 1: If there is no path from Y to X in $G_{\underline{X}}$, then there cannot be a path from any of Y's ancestors to X in $G_{\underline{X}}$, since any path from X to Z would be part of a path from X to Y.

Condition 2: If there is no directed path from Y to X, then there cannot be a Z that lies along a directed path from Y to X, and the lemma is trivially true.

Condition 3: If there is a set B that blocks all back-door paths from X to Y, then any back-door path from X to Z will also be a back-door path from X to Y. B must block this back-door path from X and Y. If B blocks the path between X and Z, then B also blocks the back-door path from X to Z, and the query $P(z|\hat{x})$ is identifiable. If B blocks the path between Z and Y, then we can use the fact that the query $P(b|\hat{x})$ must be identifiable. If $P(b|\hat{x})$ is identifiable by Condition 4, then $P(z|\hat{x})$ must also be identifiable by Condition 4, since the variables that meet the specifications for Z_1 in condition 4 for $P(b|\hat{x})$ will also meet the specifications for Z_1 in Condition 4 for $P(z|\hat{x})$. If $P(b|\hat{x})$ is identifiable by Condition 3, then there is some B' that blocks the back-door path from X to B; it must be either between X and Z, in which case $P(z|\hat{x})$ is identifiable, or it must be between Z and B'. Since there are a finite number of links between Z and Y, eventually the back-door path from X to Z must be blocked, and the query $P(z|\hat{x})$ is identifiable.

Condition 4: If there exists a set Z_1 and Z_2, Z can come either before or after Z_1. If it comes after Z_1, then the conditions that held for Y will also hold for Z, and the query $P(y|\hat{x})$ will be identifiable. If it comes before Z_1, then $\{Z_1, Z_2\}$ will block all back-door paths from X to Z, and the query will also be identifiable. □

Proof (of Theorem 14):

(By contradiction). Let Y_1 and Y_2 be two subsets of nodes such that either no nodes Y_1 are descendants of X, or all nodes Y_1 and Y_2 are descendants of X and all nodes Y_1 are non descendants of Y_2. Assume that there exists a reducing sequence for both $P(y_2|\hat{x})$ and $P(y_1|\hat{x}, y_2)$, but not for $P(y_2|\hat{x}, y_1)$. There are three possible cases:

Case 1: Y_1 and Y_2 are both nondescendants of X. In this case, $P(y_1|\hat{x}, y_2) = P(y_1|y_2)$ and thus is identifiable.

Case 2: Y_2 is a descendant of X, but Y_1 is not. In this case, Y_1 must unblock a back-door path from X to Y_1 which cannot be blocked by conditioning on other variables. But if this is the case, then there must be an unblockable back-door path from X to Y_1. Since Y_1 is a descendant of X, that would make $P(y_1|\hat{x}, y_2)$ unidentifiable.

Case 3: Y_1 and Y_2 are both descendants of X. Y_1 cannot unblock a back-door path from X to Y_2 since Y_1 is an ancestor of Y_2. Thus $P(y_2|\hat{x})$ must be unidentifiable, which means that $P(y_2|\hat{x}, y_1)$ is also unidentifiable. □

Proof (of Theorem 15):

Assume that there exists some set Z_1, which does not consist entirely of children of X, such that Z_1 blocks all directed paths from X to Y, and there also exists a set Z_2 that blocks all back-door paths from X to Z_1 in G, and all back-door paths from Z_1 to Y in $G_{\overline{X}}$. Let Z_1' be the intersection of the children of X with the ancestors of Y. Z_1' clearly blocks all directed paths from X to Y. Any back-door path from X to Z_1' must also be part of a back-door path from X to some member of Z_1, since every member of Z_1 must be either a member of Z_1' or a descendant of some member of Z_1'. Since Z_2 consists of non-descendants of

X, Z_2 must block all back-door paths from X to Z_1 between X and Z_1' – so Z_2 also blocks all back-door paths from X to Z_1'. Similarly, all back-door paths from Z_1' to Y are also part of back-door paths from Z_1 to Y, which are also blocked by Z_2. □

5.8 Complexity Analysis

Using the results of Section 5.7, we can show that the identifiability test provided by Theorem 12 can be implemented in polynomial time. We will show that each of the four conditions in Theorem 12 can be tested in polynomial time.

1. Since d-separation can be determined in time $O(V + E)$, Condition 1 can be tested in polynomial time.

2. Again, since d-separation can be determined in time $O(V + E)$, Condition 2 can be tested in polynomial time.

3. Theorem 13 allows us to test a single minimal blocking set to determine whether Condition 3 holds. Thus, we need to find a minimal blocking set between two variables. This can be done in polynomial time as follows.

   Function $BlockingSet(X, Y)$

   Input: Variables X and Y

   Output: Set B of variables that block all back-door paths between X and Y

   (a) Set $R_1 = X$ and $R_2 = \mathbf{pa}_X$

   (b) For each $r \in R_2$ that has a confounding (two-headed) link to a member of R_1, remove r from R_2, add r's parents to R_2, and add r to R_1

(c) If $R_2 \cap Y \neq \emptyset$, return FAIL

(d) Set $R_3 = Y$ and $R_4 = \mathbf{pa}_Y$

(e) For each $r \in R_4$ that has a confounding (two-headed) link to a member of R_3, remove r from R_4, add r's parents to R_4, and add r to R_3

(f) If $R_4 \cap X \neq \emptyset$, return FAIL

(g) Set $B = R_2 \cup R_4$

(h) If $(Y \not\perp\!\!\!\perp X|B)$, return FAIL

(i) For each member b of B, if $(Y \perp\!\!\!\perp X|B \setminus b)$, remove b from B

(j) If anything was removed from B in step i, go to step i

(k) Return B

4. To test Condition 4, we need to find a set of variables Z_1 and Z_2. Theorem 15 gives us a constant-time method for choosing Z_1. To find Z_2, we need only to find a blocking set that is not a descendant of X. We can do this by labeling the descendants of X "unobservable" and using the algorithm *BlockingSet* to find a minimal blocking set.

5.9 Deriving a Closed-Form Expression for Control Queries

The polynomial-time algorithm defined by Theorem 12 not only determines the identifiability of a control query but it also provides a closed-form expression for the value $P(y|\hat{x})$, in terms of the observed probability distribution, when such a closed form exists.

Function $ClosedForm(P(y|\hat{x}))$

Input: Control query of the form $P(y|\hat{x})$

Output: Either a closed-form expression for $P(y|\hat{x})$, in terms of observed variables only, or FAIL when query is not identifiable

1. If $(X \perp\!\!\!\perp Y)_{G_{\overline{X}}}$, then return $P(y)$
2. Otherwise, if $(X \perp\!\!\!\perp Y)_{G_{\underline{X}}}$, then return $P(y|x)$
3. Otherwise, let $B = BlockingSet(X, Y)$, and $Pb = ClosedForm(b|\hat{x})$; if $Pb \neq FAIL$, return $\sum_b P(y|b, x) * Pb$
4. Otherwise, Let $Z_1 = Children(X) \cap (Y \cup Ancestors(Y))$, $Z_3 = BlockingSet(X, Z_1)$, $Z_4 = BlockingSet(Z_1, Y)$, and $Z_2 = Z_3 \cup Z_4$; if $Y \notin Z_1$ and $X \notin Z_2$, return $\sum_{z_1,z_2} \sum_{x'} P(y|z_1, z_2, x')P(x'|z_2)P(z_1|x, z_2)P(z_2)$
5. Otherwise, return FAIL

This function returns either a closed-form representation of the value of the control query or, when the query is not identifiable, FAIL.

5.10 Conclusion

In this chapter, we devised a polynomial-time algorithm for determining the identifiability of control queries. If a query is identifiable, then the algorithm gives a closed-form representation of the value of the control query, in terms of the original probability distribution. Thus, we have a tractable method for assessing the ramifications of actions, given a qualitative causal diagram together with a probability distribution on a set of observed variables. In artificial intelligence, the primary attraction of this method is that it enables one agent to learn to

act by passively observing the performance of another acting agent, even in cases where the actions of the other agent are predicated on factors that are not visible to the learner. If the learner is permitted to act as well as observe, then task becomes much easier: the topology of the causal graph could then be at least partially inferred, and the effects of some previously unidentifiable actions could be determined. Immediate applications to cause-effect analysis of nonexperimental data in the social and medical sciences are discussed in [Pea95a].

CHAPTER 6

Conclusion

Causal mechanisms are an integral component in day-to-day to reasoning. If we rely solely on intuitive notions of causality, however, we can easily be trapped by fallacies and unsound reasoning. The pitfalls attendant upon the use of vague notions of causality have caused many scientists to give causal language a wide berth. We can avoid these pitfalls by giving mathematically rigorous definitions to causal notions. In this thesis, we formulate such definitions in the language of structural causal models. Once we have a formal basis for causal notions, we can examine their properties, and devise procedures for inferring one property from another. For instance, by comparing the set of axioms that arise from our definition of causal irrelevance to the set that governs path-interception in directed graphs, we can both inform our intuition and track how closely the workings of causality in our formal system match our understanding of the physical world.

The formal system of structural causal models has two major advantages. For one, it offers researchers and policy analysts a unified and unambiguous language in which to describe world models. Because causal models force all assumptions of each party to be made explicit, and the ramifications of each model can be computed easily, it is possible for researchers to discover exactly how two competing world models disagree and to judge which is more applicable to a given situation, or if yet another model is more appropriate. In some cases,

differing models can be combined, and the relative strength of various connections can be determined from data.

The other advantage is that formal systems such as structural models are conducive to mechanical computation. Thus, we can create algorithms to do reasoning and answer queries about causation using standard digital computers. This thesis has presented one such algorithm, which determines the causal effect of one variable on another from data obtained under uncontrolled conditions.

Structural causal models offer a mathematically rigorous means for specifying and examining our intuitive notions of causality, an unambiguous language describing knowledge about cause-effect relationships, and a computational device for answering causal queries.

APPENDIX A

Counterexamples

2.2.3 $(XW \not\rightarrow Y|Z)_P \Longrightarrow (X \not\rightarrow Y|Z)_P \vee (X \not\rightarrow W|Z)_P$.

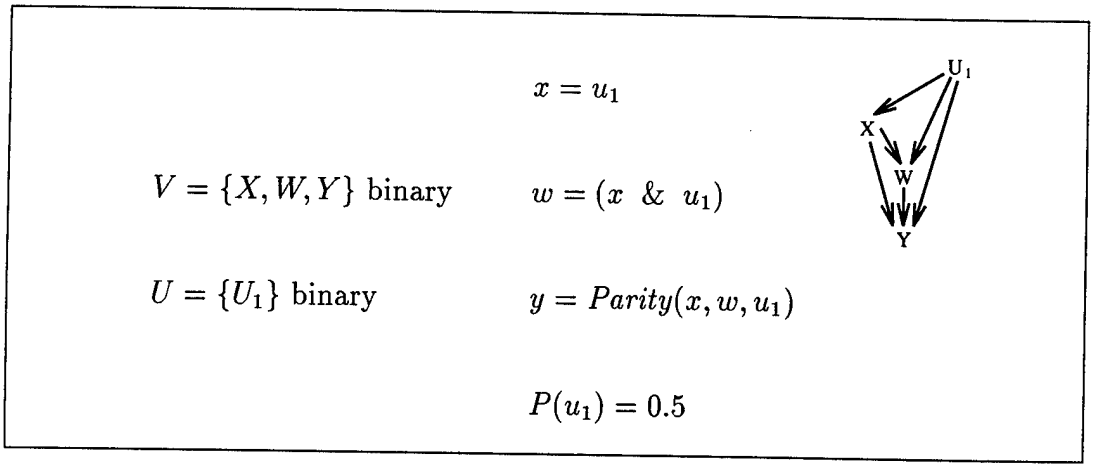


Figure A.1: Counterexample to property 2.2.3

In the causal model of Figure A.1, we can see that $(XW \not\rightarrow Y|\emptyset)_P$ & $\neg(X \not\rightarrow W|\emptyset)_P$ & $\neg(X \not\rightarrow Y|\emptyset)_P$.

In this counterexample, changing X can affect the probability of Y, and changing X can affect the probability of W, but changing X and W together cannot affect the probability of Y. Since changing X affects the value of W, it makes sense to think that intervening on W while intervening on X would not interfere with the effect that X has on Y. However, X does not completely control W. That is, when we only intervene on X, U_1 still has

some effect on W. Controlling both X and Y removes the influence of U_1 on W. As in the counterexample to property 2.2.2, removing the connection between U_1 and W prevents X from having an effect on Y.

2.2.4 $(XW \not\rightarrow Y|Z)_P \;\&\; (XY \not\rightarrow W|Z)_P \Longrightarrow (X \not\rightarrow Y|Z)_P \vee (X \not\rightarrow W|Z)_P$.

$$x = u_1$$

$$V = \{X, W, Y\} \text{ binary} \qquad y = \begin{cases} u_1 & \text{if } u_2 = 0 \\ Parity(x, w, u_1) & \text{if } u_2 = 1 \end{cases}$$

$$U = \{U_1, U_2\} \text{ binary} \qquad w = \begin{cases} Parity(x, y, u_1) & \text{if } u_2 = 0 \\ u_1 & \text{if } u_2 = 1 \end{cases}$$

$$P(u_1) = P(u_2) = 0.5$$

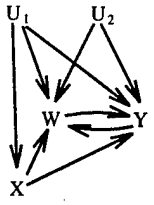


Figure A.2: Counterexample to property 2.2.4

In Figure A.2, we can see that

$P(w) = P(y) = 0.5$;

$P(w|set(X = 1)) = P(y|set(X = 1)) = 0.75$;

$P(w|\hat{x}, \hat{y}) = 0.5$ for all values of $\hat{x}, \hat{y}$; and

$P(y|\hat{x}, \hat{w}) = 0.5$ for all values of $\hat{x}, \hat{w}$

Thus, $(XW \not\rightarrow Y|\emptyset)_P \;\&\; (XY \not\rightarrow W|\emptyset)_P \;\&\; \neg((X \not\rightarrow Y|\emptyset)_P \vee (X \not\rightarrow W|\emptyset)_P)$.

This counterexample actually contains two causal models, each similar to the causal model of the counterexample to property 2.2.2. In one, W is a function of X, Y, and U_1, and Y is a function of U_1. As in the counterexample to property 2.2.2, X can affect W when Y has the same value as U_2,

but X has no effect on $P(w)$ when Y is held constant. In the other, W is a function of U_1, and Y is a function of X, W, and U_1. Also as in the counterexample to property 2.2.2, X can affect Y when W has the same value as U_1, but X has no effect on $P(w)$ when W is fixed. U_2 determines which model is in effect at any given time. While intervening only on X can affect $P(w)$ and $P(y)$, simultaneously changing X and Y has no effect on $P(w)$, and simultaneously changing X and W has no effect on $P(y)$.

2.3 $(X \not\rightarrow WY|Z)_P \Longrightarrow (X \not\rightarrow Y|ZW)_P$.

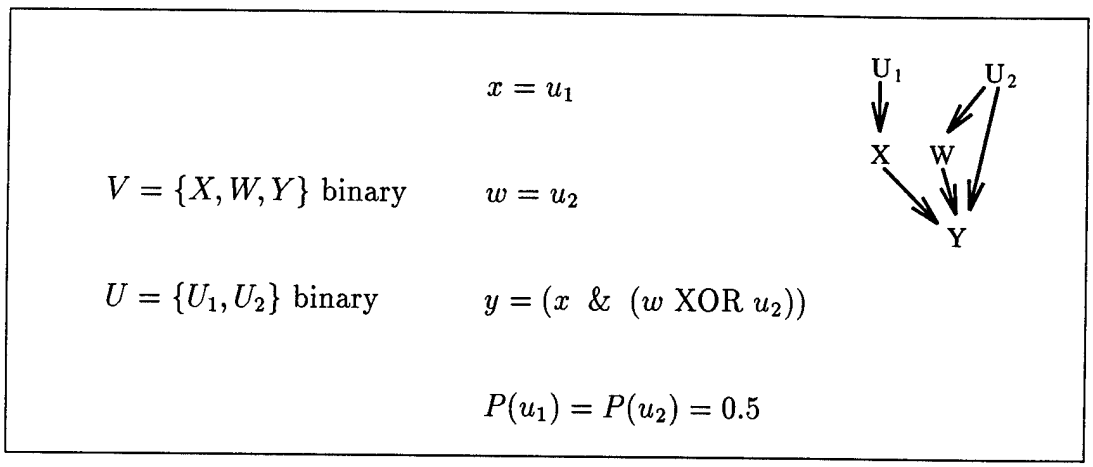


Figure A.3: Counterexample to property 2.3

In the causal model of Figure A.3, $(X \not\rightarrow YW|\emptyset)_P \ \& \ \neg(X \not\rightarrow Y|W)_P$.

In this counterexample, X does not have any effect on Y since $P(y) = 0$ and X can only act as an inhibitor of Y. When we intervene on W, then it is possible for Y to have the value 1, and X can affect the probability of Y. Thus, X can only affect Y when we intervene on W, and X has no effect on W.

2.4 $(X \not\rightarrow Y|Z)_P \ \& \ (X \not\rightarrow W|ZY)_P \Longrightarrow (X \not\rightarrow WY|Z)_P$.

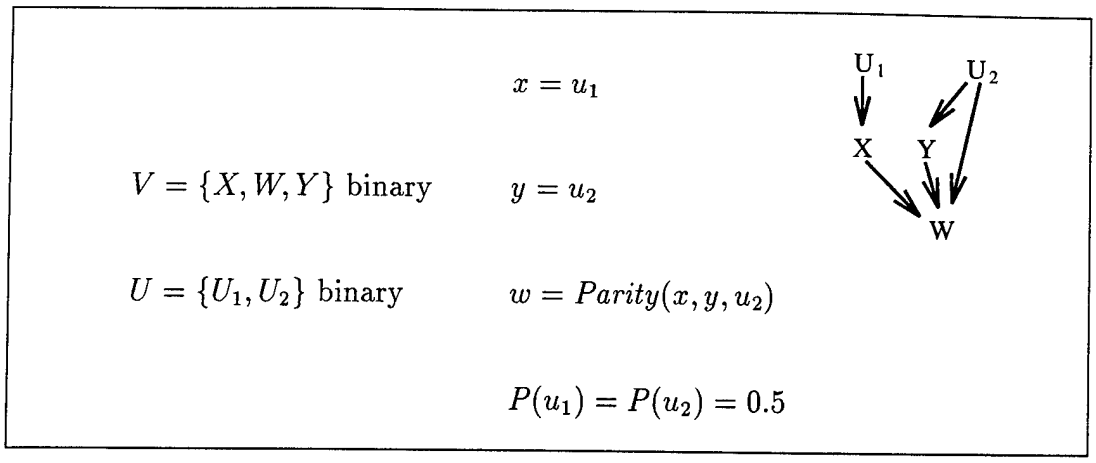


Figure A.4: Counterexample to property 2.4

In the causal model of Figure A.4, $(X \not\rightarrow Y|\emptyset)_P$ & $(X \not\rightarrow W|Y)_P$ & $\neg(X \not\rightarrow WY|\emptyset)_P$.

Changing X can affect $P(w)$ (and hence $P(y, w)$) when Y is not held fixed, and changing X has no effect on $P(y)$, but fixing Y blocks the effect that X has on W.

2.5.1 $(X \not\rightarrow Y|ZW)_P$ & $(X \not\rightarrow W|ZY)_P \Longrightarrow (X \not\rightarrow WY|Z)_P$.

In the causal model of Figure A.5, $(X \not\rightarrow Y|W)_P$ & $(X \not\rightarrow W|Y)_P$ & $\neg(X \not\rightarrow WY|\emptyset)_P$.

Fixing W prevents X from altering the probability of Y, and fixing Y prevents X from altering the probability of W, but X can change the probability of W (and hence the probability of W & Y) if there is no intervention on Y.

Up to this point, all of the counterexamples have relied on some exogenous variable from U having two different children in V. Obviously, this is not essential,

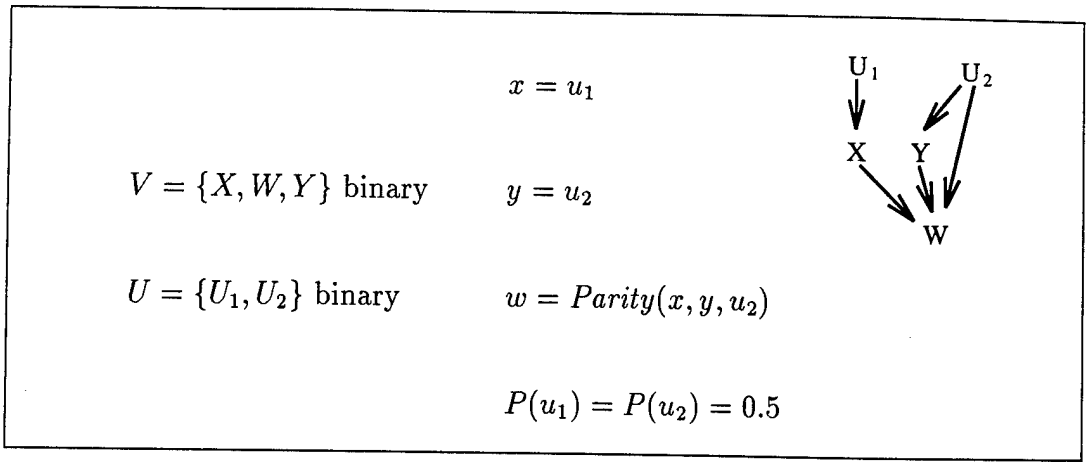


Figure A.5: Counterexample to property 2.5.1

since we could always create similar examples in which each exogenous variable has exactly one child. For instance, in the causal model of Figure A.5, we can replace U_2 with Z to get the model of Figure A.6.

In this model, all of the exogenous variables U have exactly one child, yet property 2.5.1 still does not hold. There is still an undirected cycle in the underlying causal graph, which is required for property 2.5.1 to be false. Properties 2.2.1–2.6 are all true for all causal models whose causal graphs are trees. In addition, properties 2.2.1–2.5.2 are true for all causal models whose causal graphs are polytrees. Property 2.6, as we will see now, is not always true, even when we restrict its causal graph to be a polytree.

2.6 $(X \not\to Y|Z)_P \Longrightarrow (a \not\to Z|Y)_P \vee (X \not\to a|Z)_P \quad \forall a \notin X \cup Z \cup Y.$

In the causal model of Figure A.7, $(X \not\to Y|\emptyset)_P$ & $\neg(W \not\to Y|\emptyset)_P$ & $\neg(X \not\to W|\emptyset)_P$ & $W \notin X \cup Z \cup Y$.

X can only cause a minor change in W, while a large change in W is required to affect Y. Thus, X can affect W, and W can affect Y, but X has no effect

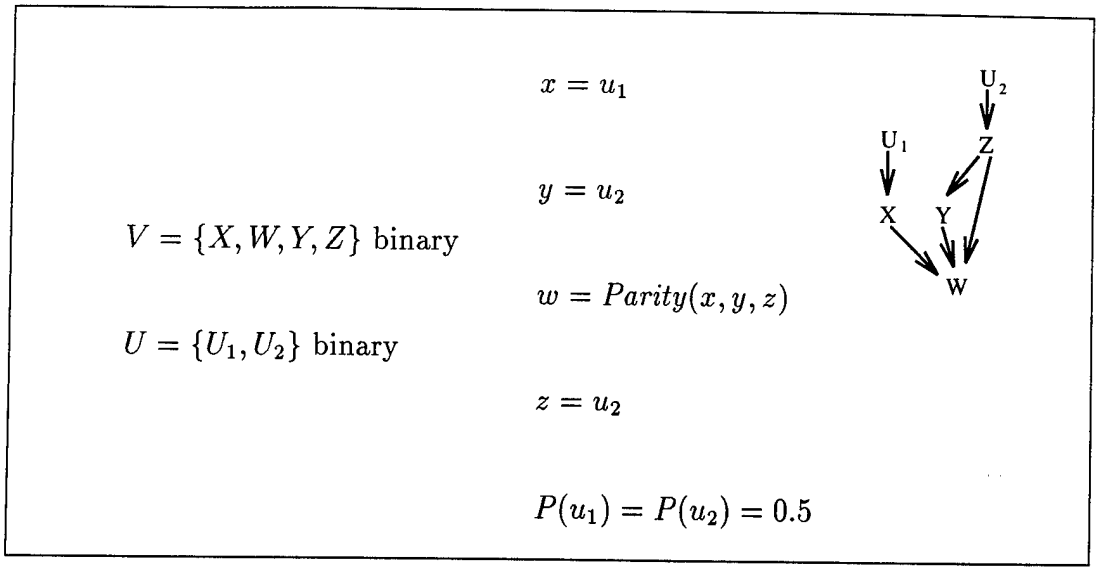


Figure A.6: Counterexample to 2.5.1, such that each variable in U has a single child

on W. Even if we restrict all variables to be binary, transitivity will not hold. For this counterexample, W could be split into four binary variables $W_1, \ldots, W_4$, such that $f_{W_1} = \neg(x \vee u_2)$, $f_{W_2} = x \ \& \ \neg u_2$, $f_{W_3} = \neg x \ \& \ u_2$, $f_{W_4} = x \ \& \ u_2$, $f_y = w_3 \vee w_4$. In Section 4.4.3, we elaborate this case.

$V = \{X, W, Y\}$,

$x = u_1$

$x, y \in \{0, 1\}$,

$w = x + 2 * u_2$

$w \in \{0, 1, 2, 3\}$

$y = (w > 1)$

$U = \{U_1, U_2\}$

$P(u_1 = 1) = P(u_2 = 1) = 0.5$

$u_1, u_2 \in \{0, 1\}$

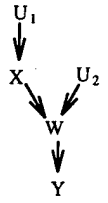


Figure A.7: Counterexample to property 2.6.

REFERENCES

[Bal95] A. Balke. *Probabilistic counterfactuals: Semantics, computation, and applications.* PhD thesis, Computer Science Department, University of California, Los Angeles, 1995.

[BP94] A. Balke and J. Pearl. "Counterfactual probabilities: Computation methods, bounds, and applications." In R.L. de Mantaras and D. Poole, editors, *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pp. 11–18, San Francisco, 1994. Morgan Kaufmann.

[BP95] A. Balke and J. Pearl. "Counterfactuals and policy analysis in structural models." In P. Besnard and S. Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 11–18, San Francisco, 1995. Morgan Kaufmann.

[Car89] N. Cartwright. *Nature's Capacities and Their Measurement.* Clarendon Press, Oxford, England, 1989.

[Daw79] A.P. Dawid. "Conditional independence in statistical theory." *Journal of the Royal Statistical Society, Series A*, **41**:1–31, 1979.

[Eel91] E. Eells. *Probabilistic Causality.* Cambridge University Press, Cambridge, England, 1991.

[FHM95] R. Fagin, J. M. Halpert, Y. Moses, and M.Y. Vardi. *Reasoning About Knowledge.* The MIT Press, Cambridge, Massachusetts, 1995.

[Fis66] F.M. Fisher. *The Identification Problem in Econometrics.* McGraw-Hill, New York, 1966.

[Fis70] F.M. Fisher. "A correspondence principle for simultaneous equation models." *Econometrica*, **38**:73–92, 1970.

[FN72] R.E. Fikes and N.J. Nilsson. "STRIPS: A new approach to the application of theorem proving to problem solving." *Artificial Intelligence*, **3**:251–284, 1972.

[Fre87] D. Freedman. "As others see us: A case study in path analysis." *Journal of Educational Statistics*, **12**:101–223, 1987. [with discussion].

[GH81] A. Gibbard and L. Harper. "Counterfactuals and two kinds of expected utility." In W.L. Harper, R. Stalnaker, and G. Pearce, editors, *Ifs.* D. Reidel, Dordrecht: Holland, 1981.

[Gol73] A.S. Goldberger. *Structural Equation Models in the Social Sciences.* Seminar Press, New York, 1973.

[Gol92] Arthur S. Goldberger. "Models of substance [comment on N. Wermuth, 'On block-recursive linear regression equations']." *Brazilian Journal of Probability and Statistics*, **6**:1–56, 1992.

[Goo61] I.J. Good. "A causal calculus." *Philosophy of Science*, **11**:305–318, 1961.

[GP92] M. Goldszmidt and J. Pearl. "Rank-based systems: A simple approach to belief revision, belief update, and reasoning about evidence and actions." In B. Nebel, C. Rich, and W. Swartout, editors, *Proceedings of the Third International Conference on Knowledge Representation and Reasoning*, pp. 661–672, San Mateo, CA, October 1992. Morgan Kaufmann Publishers.

[GP95] D. Galles and J. Pearl. "Testing identifiability of causal effects." In P. Besnard and S. Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 185–195, San Francisco, 1995. Morgan Kaufmann.

[GP97] D. Galles and J. Pearl. "An axiomatic characterization of causal counterfactuals." Technical Report R-250, Computer Science Department, University of California, Los Angeles, 1997.

[GVP90a] D. Geiger, T.S. Verma, and J. Pearl. "Identifying independence in Bayesian networks." *Networks*, **20**:507–534, 1990.

[GVP90b] D. Geiger, T.S. Verma, and J. Pearl. "Identifying Independence in Bayesian Networks." In *Networks*, volume 20, pp. 507–534. John Wiley, Sussex, England, 1990.

[Haa43] T. Haavelmo. "The statistical implications of a system of simultaneous equations." *Econometrica*, **11**:1–12, 1943.

[Hal97] J. Halpern. "Axiomatizing Causal Structures." unpublished report, Cornell University, May 1997.

[HM81] R.A. Howard and J.E. Matheson. "Influence Diagrams." *Principles and Applications of Decision Analysis, Strategic Decisions Group*, 1981.

[Hol86] P.W. Holland. "Statistics and Causal Inference [with discussion]." *Journal of the American Statistical Association*, **81**:945–970, 1986.

[HS94] D. Heckerman and R. Shachter. "A decision-based view of causality." In R. Lopez de Mantaras and D. Poole, editors, *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, 1994. Morgan Kaufmann.

[HS95] D. Heckerman and R. Shachter. "A definition and graphical representation of causality." In P. Besnard and S. Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 262–273, San Francisco, 1995. Morgan Kaufmann.

[IS86] Y. Iwasaki and H.A. Simon. "Causality in device behavior." *Artificial Intelligence*, **29**(1):3–32, 1986.

[KR51] T.C. Koopmans and O. Reiersøl. "The identification of structural characteristics." *Annals of Mathematical Statistics*, **21**:165–180, 1951.

[Lea85] E. Leamer. "Vector autoregression for causal inference?" *Carnegie-Rochester Conference Series on Public Policy*, **22**:255–304, 1985.

[Lew73a] D. Lewis. "Causation." *Journal of Philosophy*, **70**:556–567, 1973.

[Lew73b] D. Lewis. *Counterfactuals.* Harvard University Press, Cambridge, MA, 1973.

[Lew81] D. Lewis. "Counterfactuals and comparative possibility." In W.L. Harper, R. Stalnaker, and G. Pearce, editors, *Ifs.* D. Reidel, Dordrecht, Holland, 1981.

[Man90] C.F. Manski. "Nonparametric bounds on treatment effects." *American Economic Review, Papers and Proceedings*, **80**:319–323, 1990.

[Med69] J.S. Meditch. *Stochastic Optimal Linear Estimation and Control.* McGraw-Hill, New York, 1969.

[Ney23] J. Neyman. "On the application to probability theory to agricultural experiments." *Statistical Science*, **2**:465–480, 1923. Transl. (1990) from Essay on Principles, Section 9.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, CA, 1988. (Revised 4th printing, 1997).

[Pea93] J. Pearl. "Graphical models, causality, and intervention." *Statistical Science*, **8**(3):266–273, 1993.

[Pea94] J. Pearl. "A probabilistic calculus of actions." In R.L. de Mantaras and D. Poole, editors, *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pp. 454–462, San Francisco, 1994. Morgan Kaufmann.

[Pea95a] J. Pearl. "Causal Diagrams for Empirical Research [with discussion]." *Biometrika*, **82**(4):669–709, 1995.

[Pea95b] J. Pearl. "On the Testability of Causal Models with Latent and Instrumental Variables." In D. Besnard and S. Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 435–443, San Francisco, 1995. Morgan Kaufmann.

[Pea96a] J. Pearl. "Causation, action, and counterfactuals." In R. Fagin, editor, *Proceedings of the Sixth Conference Theoretical Aspects of Reasoning about Knowledge: (TARK 1996)*, pp. 51–73, San Francisco, 1996. Morgan Kaufmann.

[Pea96b] J. Pearl. "Structural and probabilistic causality." *Psychology of Learning and Motivation*, **34**:393–435, 1996.

[PP87] J. Pearl and A. Paz. "Graphoids: A graph-based logic for reasoning about relevance relations." In B. du Boulay, D. Hogg, and L. Steels, editors, *Advances in Artificial Intelligence-II*, pp. 357–363. North-Holland, Amsterdam, 1987.

[PP94] A. Paz and J. Pearl. "Axiomatic characterization of directed graphs." Technical Report R-234, Computer Science Department, University of California, Los Angeles, 1994.

[PPU96] A. Paz, J. Pearl, and S. Ur. "A new characterization of graphs based on interception relations." *Journal of Graph Theory*, **22**(2):125–136, 1996.

[PR95] J. Pearl and J. Robins. "Probabilistic evaluation of sequential plans from causal models with hidden variables." In P. Besnard and

S. Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 444–453, San Francisco, 1995. Morgan Kaufmann.

[PV91] J. Pearl and T. Verma. "A theory of inferred causation." In J.A. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the 2nd International Conference*, pp. 441–452, San Mateo, CA, 1991. Morgan Kaufmann. Also in D. Prawitz, B. Skyrms and D. Westertahl (Eds.), *Logic, Methodology and Philosophy of Science IX*, Elsevier Science B.V., 789-811, 1994.

[Rob86a] F. Robert. *Discrete Iterations, A Metric Study.* Springer-Verlag, Berlin, Germany, 1986. Trans. J. Rokne.

[Rob86b] J. Robins. "A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect." *Mathematical Modeling*, **7**:1393–512, 1986.

[Rob87] J. Robins. "Addendum to 'A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect'." *Computers and Mathematics, with Applications.*, **14**:923–45, 1987.

[RR83] P. Rosenbaum and D. Rubin. "The central role of propensity score in observational studies for causal effects." *Biometrika*, **70**:41–55, 1983.

[Rub74] D.B. Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, **66**:688–701, 1974.

[Sal84] W. Salmon. *Scientific Explanation and the Causal Structure of the World.* Princeton University Press, Princeton, 1984.

[Sav54] L.J. Savage. *The Foundations of Statistics*, volume 1. John Wiley, New York, 1954.

[SGS93] P. Spirtes, C. Glymour, and R. Schienes. *Causation, Prediction, and Search.* Springer-Verlag, New York, 1993.

[Sha96] G. Shafer. *The Art of Causal Conjecture.* MIT Press, Cambridge, MA, 1996.

[Sim70] H.A. Simon. "Causal ordering and identifiability." In W.C. Hood and T.C. Koopmans, editors, *Studies in Econometric Method*, pp. 49–74. Yale University Press, New York, [1953] 1970.

[Sob90] M.E. Sobel. "Effect analysis and causation in linear structural equation models." *Psychometrika*, **55**:495–515, 1990.

[Spo80] W. Spohn. "Stochastic independence, causal independence, and shieldability." *Journal of Philosophical Logic*, **9**:73–99, 1980.

[Stu92] M. Studeny. "Conditional independence relations have no complete characterization." In *Information Theory, Statistical Decision Functions, Random Processes: Transactions of the Eleventh Prague Conference, 1990*, pp. 377–396, Dordrecht, Holland, 1992. Kluwer Academic.

[Sup70] P. Suppes. *A Probabilistic Theory of Causation.* North-Holland, Amsterdam, 1970.

[SW60] R.H. Strotz and O.A. Wold. "Recursive versus nonrecursive systems: An attempt at synthesis." *Econometrica*, **28**:417–427, 1960.

[Wer92] N. Wermuth. "On block-recursive linear regression equations." *Brazilian Journal of Probability and Statistics*, **6**:1–56, 1992.